Nonparametric Bayesian covariate-adjusted estimation of the Youden index

Biometrics 000, 000–000 DOI: 000
000 0000

Nonparametric Bayesian Covariate-Adjusted Estimation of the Youden Index

Vanda Inácio de Carvalho[1,*], Miguel de Carvalho[1], and Adam J. Branscum[2]

[1] School of Mathematics, University of Edinburgh, Scotland, UK

[2]College of Public Health and Human Sciences, Oregon State University, USA

email: Vanda.Inacio@ed.ac.uk

Summary: A novel nonparametric regression model is developed for evaluating the covariate-specific accuracy of a continuous biological marker. Accurately screening diseased from nondiseased individuals and correctly diagnosing disease stage are critically important to health care on several fronts, including guiding recommendations about combinations of treatments and their intensities. The accuracy of a continuous medical test or biomarker varies by the cutoff threshold (c) used to infer disease status. Accuracy can be measured by the probability of testing positive for diseased individuals (the true positive probability or sensitivity, $\mathrm{Se}(c)$, of the test) and the true negative probability (specificity, $\mathrm{Sp}(c)$) of the test. A commonly used summary measure of test accuracy is the Youden index, $\mathrm{YI} = \max\{\mathrm{Se}(c) + \mathrm{Sp}(c) - 1 : c \in \mathbb{R}\}$, which is popular due in part to its ease of interpretation and relevance to population health research. In addition, clinical practitioners benefit from having an estimate of the optimal cutoff that maximizes sensitivity plus specificity available as a byproduct of estimating YI. We develop a highly flexible nonparametric model to estimate YI and its associated optimal cutoff that can respond to unanticipated skewness, multimodality and other complexities because data distributions are modeled using dependent Dirichlet process mixtures. Important theoretical results on the support properties of the model are detailed. Inferences are available for the covariate-specific Youden index and its corresponding optimal cutoff threshold. The value of our nonparametric regression model is illustrated using multiple simulation studies and data on the age-specific accuracy of glucose as a biomarker of diabetes.

Key words: Diagnostic test; Dirichlet process mixtures; Optimal cutoff; Sensitivity; Specificity.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Evaluating and ranking the performance of medical tests for screening and diagnosing disease greatly contributes to the health promotion of individuals and communities. The ability of a medical test to distinguish diseased from nondiseased individuals must be thoroughly vetted before the test can be widely used in practice. Throughout this paper we use the terms "medical test" and "test" to broadly include any continuous classifier (e.g., a single biological marker or a composite score from a combination of biomarkers) for a well-defined condition (termed "disease," with "nondiseased" used to indicate absence of the condition). The ability of a test that produces outcomes on a continuous scale to correctly differentiate between diseased (D) and nondiseased ($\bar{D}$) individuals is characterized by the separation between the distributions of test outcomes for the D and $\bar{D}$ populations. A common parametric approach to data analysis assumes that data from the D and $\bar{D}$ populations vary according to separate normal distributions. As a safeguard against model misspecification and to permit robustness from the sharp constraints of parametric models (e.g., the normal-normal model) that can fail to accommodate increasingly complex distributions of data from modern medical tests, many contemporary methods for estimating test accuracy are based on flexible statistical models that use nonparametric or semiparametric structures (e.g., Erkanli et al., 2006; Wang et al., 2007; Branscum et al., 2008; Hanson et al., 2008; Gonzalez-Manteiga et al., 2011; Inácio et al., 2011; Inácio de Carvalho et al., 2013; Rodríguez and Martínez, 2014; Zhao et al., 2015). We develop a nonparametric Bayesian regression modeling framework that allows for data-driven flexibility from the confines of parametric models by using dependent Dirichlet process mixtures to estimate the covariate-specific Youden index of a medical test and the covariate-specific optimal threshold to screen individuals in practice.

The Youden index (Youden, 1950) is a popular summary measure of the accuracy of continuous tests. Let y_D and $y_{\bar{D}}$ denote (possibly transformed) data from the D and $\bar{D}$

populations, respectively, and let F_D/f_D and $F_{\bar{D}}/f_{\bar{D}}$ denote the corresponding continuous distribution/density functions. Without loss of generality, we assume that a subject is classified as diseased (nondiseased) if the test value is greater (less) than a threshold $c \in \mathbb{R}$. Then, the probability of a positive test for a diseased subject (i.e., the sensitivity of the test) is $\mathrm{Se}(c) = \Pr(y_D > c) = 1 - F_D(c)$, and the test's specificity to correctly classify nondiseased subjects is $\mathrm{Sp}(c) = \Pr(y_{\bar{D}} \leqslant c) = F_{\bar{D}}(c)$. The Youden index (YI) is given by

$$\mathrm{YI} = \max_{c \in \mathbb{R}} \{\mathrm{Se}(c) + \mathrm{Sp}(c) - 1\} = \max_{c \in \mathbb{R}} \{F_{\bar{D}}(c) - F_D(c)\},$$

and thus combines sensitivity and specificity over all thresholds into a single numeric summary. To qualify as a bona fide medical test, it is required that $\mathrm{Se}(c) + \mathrm{Sp}(c) > 1$ for all c. Therefore, YI ranges from 0 to 1, with $\mathrm{YI} = 0$ corresponding to complete overlap of the data distributions for the D and $\bar{D}$ populations (i.e., $F_D(c) = F_{\bar{D}}(c)$ for all c) and $\mathrm{YI} = 1$ when the distributions are completely separated; values of YI between 0 and 1 correspond to different levels of stochastic ordering between $F_{\bar{D}}$ and F_D, with values closer to one indicating better discriminatory ability.

In addition to providing a global measure of test accuracy, YI provides a criterion to select an optimal threshold to screen subjects in clinical practice. The criterion is to choose the cutoff value for which sensitivity plus specificity is maximized, i.e.,

$$c^* = \arg\max_{c \in \mathbb{R}} \{F_{\bar{D}}(c) - F_D(c)\}.$$

It is worth noting that the Youden index corresponds to the maximum vertical distance between the receiver operating characteristic (ROC) curve and the chance diagonal line, with c^* being the cutoff that achieves this maximum. This criterion to select the optimal cutoff has been found to be superior to another popular approach for selecting an optimal threshold, namely using the value of c for which the ROC curve is closest to the point $(0, 1)$ in $\mathbb{R}^2$. Specifically, the ROC-based criterion can lead to an increased rate of misclassification compared to the YI-based criterion (Perkins and Schisterman, 2006). The Youden index

continues to be successfully used in practice across a variety of scientific fields (e.g., Hawass, 1997; Demir et al., 2002; Castle et al., 2003; Larner, 2015), resulting in a demand for increased research to develop flexible and robust methods that can reliably estimate it (e.g., Fluss et al., 2005; Molanes-López and Léton, 2011; Bantis et al., 2014; Zhou and Qin, 2015).

Although it is well known that the discriminatory power of a medical test is often affected by covariates, such as age or sex, past research has mostly been devoted solely to estimating the unadjusted Youden index rather than covariate-specific Youden indices and their associated optimal cutoffs. To the best of our knowledge, the only literature on estimating the covariate-specific YI has involved normal linear regression models (Faraggi, 2003), heteroscedastic kernel-based methods (Zhou and Qin, 2015), and a model-free estimation method (Xu et al., 2014). In this paper, we develop a nonparametric Bayesian regression model that is based on dependent Dirichlet process mixtures, which provide a very flexible tool that can capture a wide variety of functional forms. In contrast with most of the aforementioned models for the Youden index, where only one or two characteristics (mean and/or variance) of the distributions of test outcomes in each group depend on covariates, our modeling framework allows the entire distributions to smoothly change as a function of covariates by using B-splines regression. Therefore, our new procedure successfully combines two sources of nonparametric flexibility, namely (i) arbitrary and unspecified distributions for test outcome data from the D and $\bar{D}$ populations in place of standard parametric distributions and (ii) nonparametric regression B-splines in place of the standard linearity assumption in multiple regression.

The remainder of the paper is organized as follows. In the next section we introduce our new approach to nonparametric Bayesian estimation of the Youden index via a flexible mixture model. The performance of our methods is assessed in Section 3 using multiple simulation

studies. Section 4 applies our methods to estimate the age-specific accuracy of glucose as a biomarker of diabetes. Concluding remarks are provided in Section 5.

2. Models and Methods

We develop a nonparametric regression model to estimate the covariate-specific Youden index and optimal threshold by using dependent Dirichlet process (DDP) mixtures. The Dirichlet process (Ferguson, 1973) is a prior probability model for an unknown distribution function F and is characterized by a baseline distribution F^* (the prior mean; $\mathrm{E}\{F(\cdot)\} = F^*(\cdot)$) and a positive precision parameter α that is related to the prior variance, with larger values of α resulting in prior realizations of F that are stochastically closer to F^*. Let $F \sim \mathrm{DP}(\alpha, F^*)$ denote that F follows a Dirichlet process prior with parameters α and F^*. We will use the following constructive definition of the Dirichlet process developed by Sethuraman (1994):

$$F(\cdot) = \sum_{\ell=1}^{\infty} p_\ell \delta_{\theta_\ell}(\cdot).$$

Here, δ_θ denotes a point mass at θ, and $\theta_1, \theta_2, \ldots$ are independently distributed according to F^* and they are independent of the weights, which are generated by a stick-breaking scheme wherein $p_1 = q_1$ and for $\ell = 2, 3, \ldots$, $p_\ell = q_\ell \prod_{r=1}^{\ell-1}(1 - q_r)$, with $q_1, q_2, \ldots \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$. MacEachern (2000) proposed the DDP, a generalization of the DP, as a prior for a collection of covariate-dependent random distributions $\{F_{\mathbf{x}} : \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p\}$. Because of the full support properties it obtains (Barrientos et al., 2012), we consider a 'single-weights' DDP (De Iorio et al., 2009) in which

$$F_{\mathbf{x}}(\cdot) = \sum_{\ell=1}^{\infty} p_\ell \delta_{\theta_{\mathbf{x}\ell}}(\cdot). \tag{1}$$

The random support locations $\theta_{\mathbf{x}\ell} = \{\theta_l(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ are, for $\ell = 1, 2, \ldots$, independent and identically distributed realizations from a stochastic process over the covariate space $\mathcal{X}$ and the weights $\{p_\ell\}_{\ell=1}^{\infty}$ match those from a standard DP. We begin by describing a nonparametric model for medical test data in the absence of covariates and build up to

our new nonparametric DDP mixture model that contains nonlinear regression B-splines for capturing unforeseen complex covariate trends and that provides robust subpopulation-specific inference about YI.

2.1 *Nonparametric Model*

In the absence of covariates, we consider nonparametric data analysis using separate Dirichlet process mixture (DPM) models for data $y_{D1}, \ldots, y_{Dn_D}$ from population D and $y_{\bar{D}1}, \ldots, y_{\bar{D}n_{\bar{D}}}$ from population $\bar{D}$. That is, we consider normal mixture models with a DP prior placed on the mixing distribution, namely

$$y_{Di}|F_D \overset{\text{iid}}{\sim} F_D, \quad F_D(c) = \int \Phi(c;\mu,\sigma^2)\mathrm{d}G_D(\mu,\sigma^2), \quad G_D \sim \mathrm{DP}(\alpha_D, G_D^*),$$

where $\Phi(c;\mu,\sigma^2)$ denotes the normal distribution function with mean μ and variance σ^2 that is evaluated at c. We select the baseline distribution $G_D^*(\mu,\sigma^2)$ to be $\mathrm{N}(\mu \mid m_D, s_D^2)\Gamma(\sigma^{-2} \mid a_D, b_D)$ (i.e., $G_D^*(\mu,\sigma^2)$ is the product of independent normal and gamma distribution functions). A similar model is posited for data from the $\bar{D}$ population. The stick-breaking representation of the Dirichlet process leads to specifying the sampling models as infinite normal mixtures given by

$$F_D(c) = \sum_{\ell=1}^{\infty} p_{D\ell}\Phi(c;\mu_{D\ell},\sigma_{D\ell}^2) \quad \text{and} \quad F_{\bar{D}}(c) = \sum_{\ell=1}^{\infty} p_{\bar{D}\ell}\Phi(c;\mu_{\bar{D}\ell},\sigma_{\bar{D}\ell}^2),$$

with the aforementioned Sethuraman construction used to define the weights and priors (e.g., $\mu_{D\ell} \overset{\text{iid}}{\sim} \mathrm{N}(m_D, s_D^2)$ and $\sigma_{D\ell}^{-2} \overset{\text{iid}}{\sim} \Gamma(a_D, b_D)$). The Youden index under this nonparametric model is $\mathrm{YI} = \max_{c\in\mathbb{R}}\{\sum_{\ell=1}^{\infty}(p_{\bar{D}\ell}\Phi(c;\mu_{\bar{D}\ell},\sigma_{\bar{D}\ell}^2) - p_{D\ell}\Phi(c;\mu_{D\ell},\sigma_{D\ell}^2))\}$ and c^* is the input that returns the maximum value. Following the popular computational approach by Ishwaran and James (2002), we fit the model by accurately approximating the infinite mixtures that characterize F_D and $F_{\bar{D}}$ by finite mixtures with many components (details in Section 2.2).

2.2 *Nonparametric Regression Model*

We develop a robust nonparametric model that can be used to determine if and how the accuracy of a medical test varies across subpopulations defined by different covariate values. For ease of exposition, we assume that $p = 1$ (i.e., one covariate); an extension to the multiple covariate case is outlined at the end of this section. In this setting, sensitivity and specificity depend on a single covariate x, so that $\text{Se}(c \mid x) = \Pr(y_D > c \mid x) = 1 - F_D(c \mid x)$ and $\text{Sp}(c \mid x) = \Pr(y_{\bar{D}} \leqslant c \mid x) = F_{\bar{D}}(c \mid x)$. The data from population D are $\{(y_{Di}, x_{Di}) : i = 1, \ldots, n_D\}$ and from population $\bar{D}$ we have $\{(y_{\bar{D}j}, x_{\bar{D}j}) : j = 1, \ldots, n_{\bar{D}}\}$, where $x_{Di}, x_{\bar{D}j} \in \mathcal{X} \subseteq \mathbb{R}$ for all i, j. Test outcomes are assumed to be independent with $y_{Di} \mid x_{Di} \overset{\text{ind.}}{\sim} F_D(\cdot \mid x_{Di})$ and $y_{\bar{D}j} \mid x_{\bar{D}j} \overset{\text{ind.}}{\sim} F_{\bar{D}}(\cdot \mid x_{\bar{D}j})$. For $x \in \mathcal{X}$, the covariate-specific Youden index and optimal cutoff are given by

$$\text{YI}(x) = \max_{c \in \mathbb{R}} \{F_{\bar{D}}(c \mid x) - F_D(c \mid x)\} \quad \text{and} \quad c^*(x) = \arg\max_{c \in \mathbb{R}} \{F_{\bar{D}}(c \mid x) - F_D(c \mid x)\}. \tag{2}$$

Note that we can also estimate $\text{YI}(x_D, x_{\bar{D}})$ and $c^*(x_D, x_{\bar{D}})$, the Youden index and optimal cutoff for diseased subjects with covariate x_D and nondiseased subjects with covariate $x_{\bar{D}}$.

We specify a prior probability model for the entire collection of conditional distributions $\mathcal{F}_d = \{F_d(\cdot \mid x) : x \in \mathcal{X}\}$ for $d \in \{D, \bar{D}\}$, where the conditional distributions in each population are characterized by covariate-dependent mixtures of normals

$$F_d(c \mid x) = \int \Phi(c; \mu, \sigma^2) \mathrm{d}G_{dx}(\mu, \sigma^2), \quad d \in \{D, \bar{D}\}, \tag{3}$$

with the single weights DDP prior in (1) placed on the mixing measure $G_{dx}(\cdot)$. Specifically, we set $\theta_{d\ell}(x) = (\mu_{d\ell}(x), \sigma^2_{d\ell})$, where the potentially nonlinear function $\mu_{d\ell}(x)$ is approximated by a linear combination of cubic B-spline basis functions over a sequence of knots $\xi_{d0} < \xi_{d1} < \cdots < \xi_{dK} < \xi_{d,K+1}$. The knots ξ_{d0} and $\xi_{d,K+1}$ are boundary knots and $\xi_{d1}, \ldots, \xi_{dK}$ are interior knots. Thus,

$$\mu_{d\ell}(x) = \sum_{q=1}^{Q} \beta_{d\ell q} B_{dq}(x), \quad Q = K + 4, \tag{4}$$

where $B_{dq}(x)$ corresponds to the qth cubic B-spline basis function in group d evaluated at x. For simplicity, we have assumed the same number of interior knots for the D and $\bar{D}$ groups.

An important issue regarding the application of regression splines is the selection of interior knots, i.e., the number of inner knots and their location. As stated in Durrleman and Simon (1989), in practice often only a few knots are needed to adequately describe most of the phenomena likely to be observed in medical studies. A maximum of three or four interior knots will often suffice. The selection of K can be assisted by a model selection criterion, e.g., the log pseudo marginal likelihood (LPML) (Geisser and Eddy, 1979). We use empirical percentiles of x_d to determine knot locations. Specifically, following Rosenberg (1995), the covariate space is partitioned in accordance with the goal of having each interval containing approximately the same number of observations, which leads to setting ξ_{dk} equal to the $k/(K+1)$ percentile of x_d, for $d \in \{D, \bar{D}\}$ and $k = 1, \ldots, K$. The boundary knots are set equal to the minimum and maximum of x_d.

We proceed by noting that $\mu_{d\ell}(x)$ can be written as

$$\mu_{d\ell}(x) = \sum_{q=1}^{Q} \beta_{d\ell q} B_{dq}(x) = \mathbf{z}_d^{\mathrm{T}} \boldsymbol{\beta}_{d\ell},$$

where $\mathbf{z}_d^{\mathrm{T}} = (B_{d1}(x), \ldots, B_{dQ}(x))$ and $\boldsymbol{\beta}_{d\ell} = (\beta_{d\ell 1}, \ldots, \beta_{d\ell Q})^{\mathrm{T}}$. Thus, under this formulation, the base stochastic processes are replaced with a group-specific base distribution G_d^* that generates the component specific regression coefficients and variances. The B-splines DDP mixture model can therefore be represented as a DP mixture of Gaussian regression models where the component means vary nonlinearly with the predictors, namely

$$F_d(c \mid x) = \int \Phi(c; \mathbf{z}_d^{\mathrm{T}} \boldsymbol{\beta}, \sigma^2) \mathrm{d}G_d(\boldsymbol{\beta}, \sigma^2), \quad G_d \sim \mathrm{DP}(\alpha_d, G_d^*), \quad d \in \{D, \bar{D}\}. \tag{5}$$

To complete model (5), we take $G_d^*(\boldsymbol{\beta}, \sigma^2)$ to be $\mathrm{N}_Q(\boldsymbol{\beta} \mid \boldsymbol{m}_d, \boldsymbol{S}_d)\Gamma(\sigma^{-2} \mid a_d, b_d)$, with conjugate hyperpriors $\boldsymbol{m}_d \sim \mathrm{N}_Q(\boldsymbol{m}_{d0}, \boldsymbol{S}_{d0})$ and $\boldsymbol{S}_d^{-1} \sim \mathrm{Wishart}_Q(\nu_d, (\nu_d \boldsymbol{\psi}_d)^{-1})$ (a Wishart distribution with degrees of freedom ν_d and expectation $\boldsymbol{\psi}_d^{-1}$).

We use the blocked Gibbs sampler of Ishwaran and James (2002) for posterior sampling.

The blocked Gibbs sampler relies on truncating the stick-breaking representation to a finite number of components. Hence,

$$F_D(c \mid x) = \sum_{\ell=1}^{L_D} p_{D\ell}\Phi(c; \mathbf{z}_D^{\mathrm{T}}\boldsymbol{\beta}_{D\ell}, \sigma^2_{D\ell}) \quad \text{and} \quad F_{\bar{D}}(c \mid x) = \sum_{\ell=1}^{L_{\bar{D}}} p_{\bar{D}\ell}\Phi(c; \mathbf{z}_{\bar{D}}^{\mathrm{T}}\boldsymbol{\beta}_{\bar{D}\ell}, \sigma^2_{\bar{D}\ell}), \qquad (6)$$

with L_D and $L_{\bar{D}}$ being upper bounds on the number of components used for the approximations. The conditional distribution in each group is then estimated by a finite mixture of Gaussian regression models with the mixing weights automatically determined by the data. The weights $p_{d\ell}$ are generated from the stick-breaking representation, while $\boldsymbol{\beta}_{d\ell} \overset{\text{iid}}{\sim} \mathrm{N}_Q(\boldsymbol{m}_d, \boldsymbol{S}_d)$ and $\sigma_{d\ell}^{-2} \overset{\text{iid}}{\sim} \Gamma(a_d, b_d)$. The full conditional distributions have the conjugate forms detailed in Appendix A of the Supplementary Materials. The level of truncation can be guided by properties of $U_d = \sum_{\ell=L_d+1}^{\infty} p_{d\ell}$. Ishwaran and Zarepour (2000) demonstrated that $\mathrm{E}(U_d) = \alpha_d^{L_d}/(1+\alpha_d)^{L_d}$ and $\mathrm{var}(U_d) = \alpha_d^{L_d}/(2+\alpha_d)^{L_d} - \alpha_d^{2L_d}/(1+\alpha_d)^{2L_d}$. For example, setting $L_d = 20$ and $\alpha_d = 1$ ($d \in \{D, \bar{D}\}$) as in our simulation study and application, results in $\mathrm{E}(U_d) = 10^{-6}$ and $\mathrm{var}(U_d) \doteq 10^{-10}$, which is more than adequate for our data analysis.

Posterior inference for $\mathrm{YI}(x)$ is obtained by using (2) and (6), and the covariate-specific optimal cutoff $c^*(x)$ is the input that returns the maximum. A grid search was used to identify the maximum.

Finally, in the case of multiple covariates (so that $\mathbf{x} \in \mathbb{R}^p$), a possibility would be to use the additive structure $\mu_{d\ell}(\mathbf{x}) = f_{d\ell 1}(x_1) + \cdots + f_{d\ell p}(x_p)$. The model can therefore be regarded as a DDP mixture of additive models. Each function could be approximated by basis functions as in (4).

2.3 *Theoretical Properties*

In this section we characterize the support properties of our nonparametric models, with and without covariates. The overarching goal is to construct extremely flexible models for F_D and $F_{\bar{D}}$ that support any collection of Youden indexes with positive probability. Roughly speaking, the following results are an affirmation of the theoretical resilience of the model, in

the sense that the model can successfully adapt to and support very complex distributions of data. We have the following theorem about the nonparametric model in Section 2.1.

THEOREM 1: *Let* $(\Omega, \mathcal{A}, P)$ *be the probability space associated with the DPM model, which induces the Youden index* $YI = \max_{c\in\mathbb{R}}\{F_{\bar{D}}(c) - F_D(c)\}$. *For almost every* $\omega \in \Omega$, *let* YI_ω *be a realization of the Youden index under the proposed DPM. Then, for every* $\varepsilon > 0$, *it holds that* $P(\omega \in \Omega : |YI_\omega - YI| < \varepsilon) > 0$.

The following analogous result holds for the covariate-dependent nonparametric regression setting in Section 2.2.

THEOREM 2: *Let* $(\Omega, \mathcal{A}, P)$ *be the probability space associated with the general DDP mixture of Gaussian distributions in* (3) *with the single weights DDP prior in* (1) *placed on the mixing measure and with trajectories given by* $YI(\mathbf{x}) = \max_{c\in\mathbb{R}}\{F_{\bar{D}}(c \mid \mathbf{x}) - F_D(c \mid \mathbf{x})\}$. *For almost every* $\omega \in \Omega$ *and every* $\mathbf{x} \in \mathcal{X}$, *let* $YI_\omega(\mathbf{x})$ *be a trajectory of the Youden index under the DDP mixture model. Then, for* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, *for every positive integer* n *and* $\varepsilon > 0$, *it holds that* $P(\omega \in \Omega : |YI_\omega(\mathbf{x}_i) - YI(\mathbf{x}_i)| < \varepsilon,\ i = 1, \ldots, n) > 0$.

Proofs are given in Appendix B of the Supplementary Materials.

3. Simulation Study

To evaluate the performance of our nonparametric regression model for estimating the covariate-specific Youden index and optimal cutoff value, we analyzed simulated data under the following four scenarios: linear mean, a mixture of linear means, nonlinear mean with constant variance, and nonlinear mean with covariate-dependent variance. For each scenario, 100 data sets were generated using sample sizes of $(n_D, n_{\bar{D}}) = (100, 100)$, $(n_D, n_{\bar{D}}) = (100, 200)$, and $(n_D, n_{\bar{D}}) = (200, 200)$. For all scenarios, covariate values were independently generated from a uniform distribution, namely $x_{Di} \sim \mathrm{U}(0, 1)$ and $x_{\bar{D}j} \sim \mathrm{U}(0, 1)$.

3.1 *Simulation Scenarios*

In Scenario 1, we consider different homoscedastic linear mean regression models for the diseased and nondiseased groups. Specifically, independent data were generated as

$$y_{Di} \mid x_{Di} \sim \mathrm{N}(2{+}4x_{Di}, 2^2), \quad y_{\bar{D}j} \mid x_{\bar{D}j} \sim \mathrm{N}(0.5{+}x_{\bar{D}j}, 1.5^2), \quad i = 1, \ldots, n_D, \quad j = 1, \ldots, n_{\bar{D}}.$$

The primary purpose of including this scenario is to investigate the loss of efficiency of our covariate-specific Youden index and optimal cutoff estimators when standard parametric assumptions hold.

The popular normal-normal regression model for data from the D and $\bar{D}$ populations is violated in Scenarios 2-4. Data for Scenario 2 are governed by the following mixtures of homoscedastic linear mean regression models:

$$y_{Di} \mid x_{Di} \overset{\text{ind.}}{\sim} 0.5\mathrm{N}(2 + 3x_{Di}, 1^2) + 0.5\mathrm{N}(6 + 2.5x_{Di}, 1^2),$$

$$y_{\bar{D}j} \mid x_{\bar{D}j} \overset{\text{ind.}}{\sim} 0.5\mathrm{N}(2 + x_{\bar{D}j}, 1.25^2) + 0.5\mathrm{N}(-2.5 + x_{\bar{D}j}, 1^2).$$

Scenario 3 involves the homoscedastic nonlinear mean regression models given by

$$y_{Di} \mid x_{Di} \overset{\text{ind.}}{\sim} \mathrm{N}(9{+}1.15x_{Di}^2, 2.5^2) \quad \text{and} \quad y_{\bar{D}j} \mid x_{\bar{D}j} \overset{\text{ind.}}{\sim} \mathrm{N}(5.5{+}1.75x_{\bar{D}j}^2{+}1.5\sin(\pi(x_{\bar{D}j}{+}1)), 1.5^2),$$

for $i = 1, \ldots, n_D$, and $j = 1, \ldots, n_{\bar{D}}$. Finally, in Scenario 4, the most complex scenario considered, we use the following heteroscedastic nonlinear mean regression models for the diseased and nondiseased groups:

$$y_{Di} \mid x_{Di} \overset{\text{ind.}}{\sim} \mathrm{N}(5 + 1.5x_{Di} + 1.5\sin(x_{Di}), 1.5 + \Phi(10x_{Di} - 2)),$$

$$y_{\bar{D}j} \mid x_{\bar{D}j} \overset{\text{ind.}}{\sim} \mathrm{N}(3 + 1.5\sin(\pi x_{\bar{D}j}), 0.2 + \exp(x_{\bar{D}j})).$$

3.2 *Models*

For each simulated data set we fit the B-splines DDP mixture model with $Q = 4$, thus corresponding to $K = 0$ (no interior knots). We set α_d ($d \in \{D, \bar{D}\}$) equal to one, which according to Hanson (2006) is the default value in the absence of prior information on the number of components needed to adequately describe $F_d(\cdot \mid x)$. Using results from Antoniak

(1974) and Escobar (1994), this choice leads to a prior expected number of components of 5 when $n_d = 100$ and 6 when $n_d = 200$. For the normal-gamma prior we set $\boldsymbol{m}_{d0} = \mathbf{0}_Q, \boldsymbol{S}_{d0} = 100\mathbf{I}_Q$, $\nu_d = Q + 2$, $\boldsymbol{\psi}_d^{-1} = \mathbf{I}_Q$, where $\mathbf{I}_Q$ denotes the $Q \times Q$ identity matrix, and we used $a_d = b_d = 0.1$. The normal prior for $\boldsymbol{\beta}_{d\ell}$ is relatively diffuse since variances in $\boldsymbol{S}_{d0}$ are large and the degree of freedom in the Wishart prior is small. Although the gamma prior for $\sigma^2_{d\ell}$ has a peak at 0^+, we found that estimates of the Youden index and optimal cutoff were robust to moderate departures from this prior distribution. We capped the maximum number of mixture components at $L_d = 20$ and, thus, a maximum of 20 regression models was used to approximate the conditional distributions in (5).

We compared our estimator to results from ordinary linear regression analysis, i.e., where

$$F_D(c \mid \mathbf{x}^*) = \Phi(c; \mathbf{x}^{*\mathrm{T}}\boldsymbol{\beta}^*_D, \sigma^{2*}_D) \quad \text{and} \quad F_{\bar{D}}(c \mid \mathbf{x}^*) = \Phi(c; \mathbf{x}^{*\mathrm{T}}\boldsymbol{\beta}^*_{\bar{D}}, \sigma^{2*}_{\bar{D}})$$

with $\mathbf{x}^{*\mathrm{T}} = (1, x)$ and $\boldsymbol{\beta}^*_d = (\beta^*_{d0}, \beta^*_{d1})^{\mathrm{T}}$, $d \in \{D, \bar{D}\}$. We used the following priors that align with those from the nonparametric analysis:

$$\boldsymbol{\beta}^*_d \sim \mathrm{N}_2(\boldsymbol{m}^*_d, \boldsymbol{S}^*_d), \quad \sigma_d^{-2*} \sim \Gamma(a^*_d, b^*_d), \quad \boldsymbol{m}^*_d \sim \mathrm{N}_2(\boldsymbol{m}^*_{d0}, \boldsymbol{S}^*_{d0}), \quad \boldsymbol{S}_d^{*-1} \sim \mathrm{Wishart}(\nu^*_d, (\nu^*_d\boldsymbol{\psi}^*_d)^{-1}),$$

with

$$\boldsymbol{m}^*_{d0} = \mathbf{0}_2, \quad \boldsymbol{S}^*_{d0} = 100\mathbf{I}_2, \quad \nu^*_d = 4, \quad (\boldsymbol{\psi}^*_d)^{-1} = \mathbf{I}_2, \quad \text{and} \quad a^*_d = b^*_d = 0.1.$$

This model can be regarded as a Bayesian version of the model considered by Faraggi (2003). In both cases 1000 samples were kept after a burn-in period of 500 iterations of the Gibbs sampler. In addition, we compared our model to the nonparametric kernel-based method of Zhou and Qin (2015) (details appear in Section C of the Supplementary Materials).

3.3 *Results*

The estimated covariate-specific Youden index and optimal cutoff functions along with the 2.5 and 97.5 percentiles in Figures 1 and 2 illustrate the ability of our model to accurately and precisely capture complex functional forms dynamically. Specifically, for the case $(n_D, n_{\bar{D}}) =$

$(100, 200)$, which is similar to the diabetes application in Section 4, the minor loss in efficiency for our nonparametric estimator when a parametric normal linear regression holds (panels (a) and (e) in Figures 1 and 2) is a small price to pay for the benefit of the extreme robustness that leads to accurate data-driven estimates under increasingly complex scenarios (remaining panels in Figures 1 and 2). Similar conclusions were found for the other sample sizes (Figures 1–8 in Appendix C of the Supplementary Materials). As indicated by these figures, our estimator is able to successfully recover the true functional form of both the Youden index and optimal cutoff for all scenarios considered. As expected, the estimator based on the normality assumption has the best performance in Scenario 1, but it is unsuitable for the remaining scenarios and, unlike our nonparametric estimator, its performance fails to improve as the sample size increases. It is noteworthy that, in all scenarios considered, posterior uncertainty decreases as sample size increases and that even with a relatively low sample size combination of $(n_D, n_{\bar{D}}) = (100, 100)$, our method performs very well. The kernel-based estimator is also able to successfully recover the true functional form of the Youden index and optimal cutoff.

The discrepancy between estimated and true Youden index and optimal cutoff was assessed using the empirical global mean squared error (EGMSE)

$$\text{EGMSE}_{\text{YI}} = \frac{1}{n_x} \sum_{r=1}^{n_x} \{\widehat{\text{YI}}(x_r) - \text{YI}(x_r)\}^2,$$

where $n_x = 11$ and the x_r's are evenly spaced over $[0, 1]$. An analogous expression follows for EGMSE_{c^*}. Boxplots summarizing the distribution of EGMSE are presented in Figures 9 and 10 of the Supplementary Materials. For most sample sizes considered under these scenarios, our B-splines DDP estimator produced smaller EGMSE compared to the kernel estimator (especially for the optimal cutoff).

The mean LPML values and 90% intervals presented in Table 1 of the Supplementary Materials highlight the performance of our model compared to the normal regression model.

The conditional predictive ordinate (CPO) was calculated using M post burn-in Gibbs sampler iterates (each of which is indexed by the superscript (k) in the following formula for CPO):

$$\text{LPML}_D = \sum_{i=1}^{n_D} \log(\text{CPO}_{Di}), \qquad \text{CPO}_{Di}^{-1} = \frac{1}{M} \sum_{k=1}^{M} \left\{ \sum_{\ell=1}^{L_D} p_{D\ell}^{(k)} \phi \left(y_{Di}; \mathbf{z}_{Di}^{\mathrm{T}} \boldsymbol{\beta}_{D\ell}^{(k)}, (\sigma_{D\ell}^{(k)})^2 \right) \right\}.$$

An analogous formula holds for $\text{LPML}_{\bar{D}}$.

As suggested by a referee, we also fit the B-splines DDP model using multiple interior knots ($Q = 7$). The estimated Youden index and optimal cutoff functions are shown in Figures 1–8 of the Supplementary Materials. The true functional form is recovered successfully for both $\text{YI}(x)$ and $c^*(x)$, although with higher posterior uncertainty than with $Q = 4$. The model with $Q = 4$ was clearly favored by LPML for the majority of the scenarios and sample sizes considered (results not shown).

[Figure 1 about here.]

[Figure 2 about here.]

4. Application

Diabetes mellitus, a chronic disease characterized in part by high levels of blood sugar (glucose), is an increasingly serious global health concern, with the estimated worldwide prevalence of 8% expected to continue to rise (Shi and Hu, 2014). A population-based survey of diabetes in Cairo, Egypt collected data on postprandial blood glucose measurements that were obtained from a finger stick on 286 adults. Our primary goal is to evaluate the age-specific accuracy of glucose to serve as a biomarker of diabetes. Based on the World Health Organization diagnostic criteria for diabetes, 88 subjects were classified as diabetic and 198 as nondiabetic (Smith and Thompson, 1996).

Density estimates from an unadjusted analysis of data from the diabetic and nondiabetic groups using histograms of glucose levels and the DPM mixture of normal models described

in Section 2.1 are presented in Figures 11 (a) and (b) of the Supplementary Materials; Figure 11 (c) presents estimates of the distribution functions and optimal cutoff value. The Bayesian nonparametric estimate of the unadjusted Youden index (95% probability interval) of 0.66 $(0.56, 0.75)$ illustrates the reasonably strong overall discriminatory ability of glucose to correctly classify diabetes status. The optimal cutoff that maximizes test accuracy occurs at a glucose level of 127 mg/dL $(118, 142)$.

The aging process may be associated with relative insulin resistance among those who are nondiabetic (Smith and Thompson, 1996). Thus, there is a need to accurately estimate the Youden index and optimal cutoff value adjusted for age. Our B-splines DDP mixture model was applied to the glucose data with $Q = 4, 5, 6,$ and 7 and the same diffuse prior specification described in Section 3.2. Posterior inference was based on estimates calculated from 3500 Gibbs sampler iterates after a burn-in of the first 1500 realizations was discarded. Glucose levels were scaled by dividing by the standard deviation to fit the model, but we transformed back to the original scale to present the results (hyperparameter specification was made on the scaled data). Plots of the B-splines basis functions in each group are presented in Figure 12 of the Supplementary Materials.

Figure 3 presents the posterior mean of the Youden index and the optimal cutoff for the different values of Q as a function of age, along with a band constructed using the pointwise 2.5% and 97.5% posterior quantiles. To enable comparisons across Q, Figure 13 of the Supplementary Materials shows the 4 posterior means together on the same graph. Our analysis found that glucose is a more accurate biomarker of diabetes in younger adult populations, with its accuracy decreasing with age and the optimal cutoff increasing with age. Also, as expected, posterior uncertainty increases with Q. The LPML values for models applied to the nondiabetic data are -878, -880, -881, and -887 for $Q = 4, 5, 6,$ and 7,

respectively, while for the diabetic population the corresponding LPML values are −530, −532, −531, and −532.

Age-specific nonparametric estimates of the Youden index and optimal cutoff (posterior mean and 95% probability bands) when $Q = 4$ along with unadjusted estimates of YI and c^* are presented in Figures 4 (a) and (b). The probability band for the optimal cutoff from the unadjusted analysis is not completely contained in the age-adjusted band, which gives some additional support for estimating the age-specific accuracy of glucose for diagnosing diabetes. We compared estimates from our B-splines DDP model to those from the Gaussian linear regression model in Section 3.2 (Figures 4 (c) and (d)). To facilitate comparisons, Figures 4 (e) and (f) present the estimates together for both methods. While the estimates of the Youden index are similar, the different methods provide different estimates of the optimal age-specific cutoff values. For the nondiabetic group, the B-splines DDP model has LPML equal to -878 compared to -935 under the normal model, while for the diabetic group the corresponding LPML values are -530 and -532. The pseudo Bayes factors, which are larger than 10^{20} and 7, support the nonparametric model for both groups, although just slightly for the diabetic group. A log transformation made the normal model slightly more competitive, although the B-splines DDP mixture model was still preferred in the nondiseased group; for the diabetic group the comparison was roughly unchanged. We also highlight the important finding that our nonparametric analysis did not produce a substantial increase in the uncertainty associated with the estimates of the Youden index and optimal cutoff in this setting. A comparison with the kernel approach is presented in Figure 14 in the Supplementary Materials. A sensitivity analysis with a data-driven prior (Section D and Figure 15 of the Supplementary Materials) resulted in similar inferences as the primary analysis.

[Figure 3 about here.]

[Figure 4 about here.]

5. Concluding Remarks

We developed a Bayesian nonparametric regression model to estimate the covariate-specific Youden index and the corresponding optimal cutoff value. The flexibility of our model arises from using dependent Dirichlet process mixtures combined with B-splines regression. Our simulation study illustrated the ability of the model to dynamically respond to complex data distributions in a variety of scenarios, with little price to be paid in terms of decreased posterior precision for the extra generality of our nonparametric estimator when compared with parametric estimates (even when the parametric model holds). Our investigation into the potential of glucose to serve as a biomarker of diabetes found that its classification accuracy decreases with age and the optimal cutoff to screen subjects in practice increases with age. It is important to underscore that, although the Youden index gives equal weight to sensitivity and specificity, a weighted Youden index can also be used. For instance, weighting by the prevalence of disease in the population would emphasize test sensitivity over specificity when the disease is common. An interesting avenue for future research is variable selection in the diseased and nondiseased subpopulations; spike and slab priors could be a possible approach to this problem.

6. Supplementary Materials

Supplementary Materials describing the Gibbs sampling algorithm for fitting the nonparametric regression model, proofs of Theorems 1 and 2, details on the nonparametric kernel method of Zhou and Qin (2015), and the additional figures referenced in Sections 3 and 4 are available at the Biometrics website on Wiley Online Library.

Acknowledgements

We thank a referee for suggestions that greatly improved the paper. This research was partially funded by CONICYT, through the Fondecyt projects 11130541 and 11121186.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems *Annals of Statistics* **2**, 1152–1174.

Bantis, L. E., Nakas, C. T., and Reiser, B. (2014). Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. *Biometrics* **70**, 212–223.

Barrientos, A. F., Jara, A., and Quintana, F. (2012) On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis* **7**, 277–310.

Branscum, A.J., Johnson, W.O., Hanson, T.E., and Gardner, I.A. (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine* **27**, 2474–2496.

Castle, P., Lorincz, A. T., Scott, D. R., Sherman, M. E., Glass, A. G., Rush, B. B. *et al.* (2003). Comparison between prototype hybrid capture 3 and hybrid capture 2 human papillomavirus DNA assays for detection of high-grade cervical inter epithelial neoplasia and cancer. *Journal of Clinical Microbiology* **9**, 4022–4030.

Demir, A., Yaral, N., Fisgir, T., Duru, F., and Kara, A. (2002). Most reliable indices in differentiation between thalassemial trait and iron deficiency anemia. *Pedriatics International* **44**, 612–616.

De Iorio, M., Johnson, W. O., Müller, P. and Rosner, G. L. (2009). Bayesian nonparametric non-proportional hazards survival modelling. *Biometrics* **65**, 762–771.

Durrleman, S. and Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine* **8**, 551–561.

Erkanli, A., Sung, M., Costello, E. J., and Angold, A. (2006). Bayesian semi-parametric ROC analysis. *Statistics in Medicine* **25**, 3905–3928.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *Journal of the Royal Statistical Society, Ser. D* **52**, 1152–1174.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal* **47**, 458–472.

Geisser, S., and Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

Gonzalez-Manteiga, W., Pardo-Fernandez, J. C., and Van Keilegom, I. (2011). ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics* **38**, 169–184.

Hanson, T. E. (2006). Modeling censored lifetime data using a mixture of gammas baseline. *Bayesian Analysis* **1**, 575–594.

Hanson, T. E., Branscum, A. J., and Gardner, I. A. (2008). Multivariate mixtures of Polya trees for modelling ROC data. *Statistical Modelling* **8**, 81–96.

Hawass, N. E. D. (1997). Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. *The British Journal of Radiology* **70**, 360–366.

Inácio, V., Turkman, A. A., Nakas, C. T., and Alonzo, T. A. (2011). Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal* **53**, 1011–1024.

Inácio de Carvalho, V., Jara, A., Hanson, T. E., and de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling. *Bayesian Analysis* **8**, 623–646.

Ishwaran, H., and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.

Ishwaran, H., and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371–390.

Larner, A. J. (2015). *Diagnostic Test Accuracy Studies in Dementia: A Pragmatic Approach.* Cham: Springer.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical Report, Department of Statistics, The Ohio State University.

Molanes-López, E. M., and Letón, E. (2011). Inference of the Youden index and associated threshold using empirical likelihood quantiles. *Statistics in Medicine* **30**, 2467–2480.

Perkins, N. J., and Schisterman, E. J. (2006). The inconsistency of "optimal" cut-points using two ROC based criteria. *American Journal of Epidemiology* **163**, 670–675.

Rodríguez, A., and Martínez, J. C. (2014). Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics* **2**, 353–369.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics* **51**, 874–887.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statistica Sinica* **2**, 639–650.

Schisterman, E. F., and Perkins, N. (2007). Confidence intervals for the Youden Index and corresponding optimal cut-point. *Communications in Statistics: Simulation and Computation* **36**, 549–563.

Shi, Y., and Hu, F.B. (2014). The global implications of diabetes and cancer. *The Lancet* **383**, 1947–1948.

Smith, P. J., and Thompson, T. J. (1996). Correcting for confounding in analyzing receiver operating characteristic curves. *Biometrical Journal* **7**, 857–863.

Wang, C., Turnbull, B. W., Gröhn, Y. T., and Nielsen, S. S. (2007). Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. *Journal of Agricultural, Biological, and Environmental Statistics* **12**, 128–146.

Xu, T., Wang, J., and Fang, Y. (2014). A model-free estimation for the covariate-adjusted Youden index and its associated cut-point. *Statistics in Medicine* **33**, 4963–4974.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* **3**, 32–35.

Zhao, L., Feng, D., Chen, G., and Taylor, J. M. G. (2005). A unified Bayesian semiparametric approach to assess discrimination ability in survival analysis. *Biometrics* **72**, 554–562.

Zhou, H., and Qin, G. (2015). Nonparametric covariate adjustment for the Youden index. In *Applied Statistics in Biomedicine and Clinical Trials Design*, 109–132, Cham: Springer.

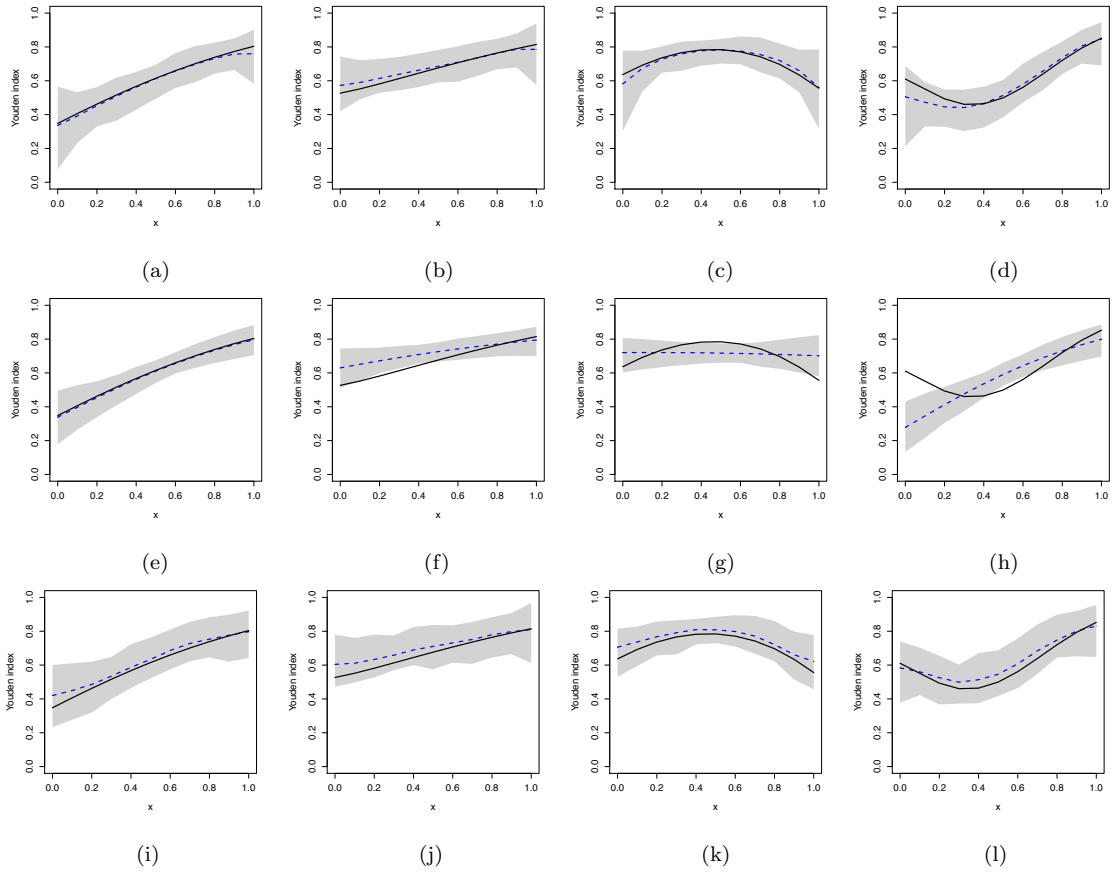


Figure 1: True (solid black lines) and the average value over 100 simulated data sets (dashed blue lines) of the posterior mean (for the Bayesian estimators) of the Youden index function for the sample size $(n_D, n_{\bar{D}}) = (100, 200)$. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator. Row 2: Normal estimator. Row 3: Kernel estimator . Panels (a), (e), and (i) show the results under Scenario 1, panels (b), (f), and (j) under Scenario 2, panels (c), (g), and (k) under Scenario 3, and panels (d), (h), and (l) under Scenario 4.

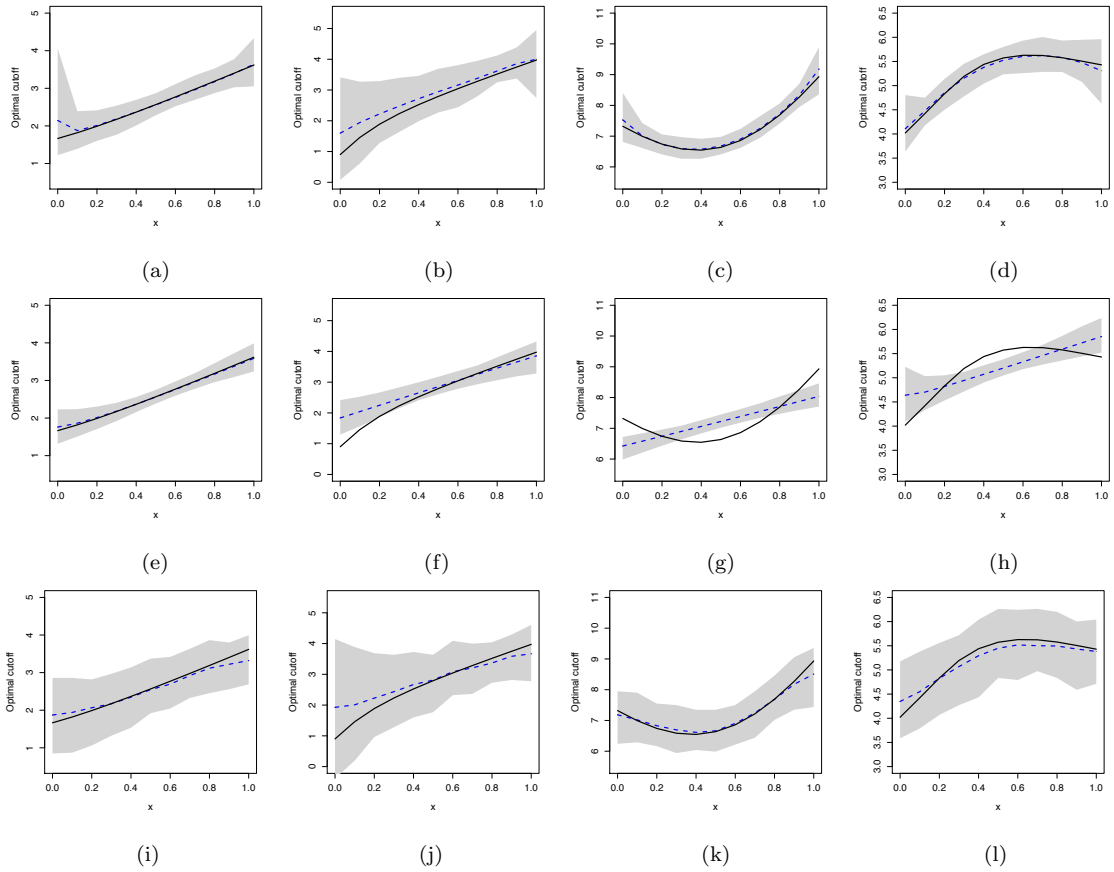


Figure 2: True (solid black lines) and the average value over 100 simulated data sets (dashed blue lines) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function for the sample size $(n_D, n_{\bar{D}}) = (100, 200)$. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator. Row 2: Normal estimator. Row 3: Kernel estimator . Panels (a), (e), and (i) show the results under Scenario 1, panels (b), (f), and (j) under Scenario 2, panels (c), (g), and (k) under Scenario 3, and panels (d), (h), and (l) under Scenario 4.

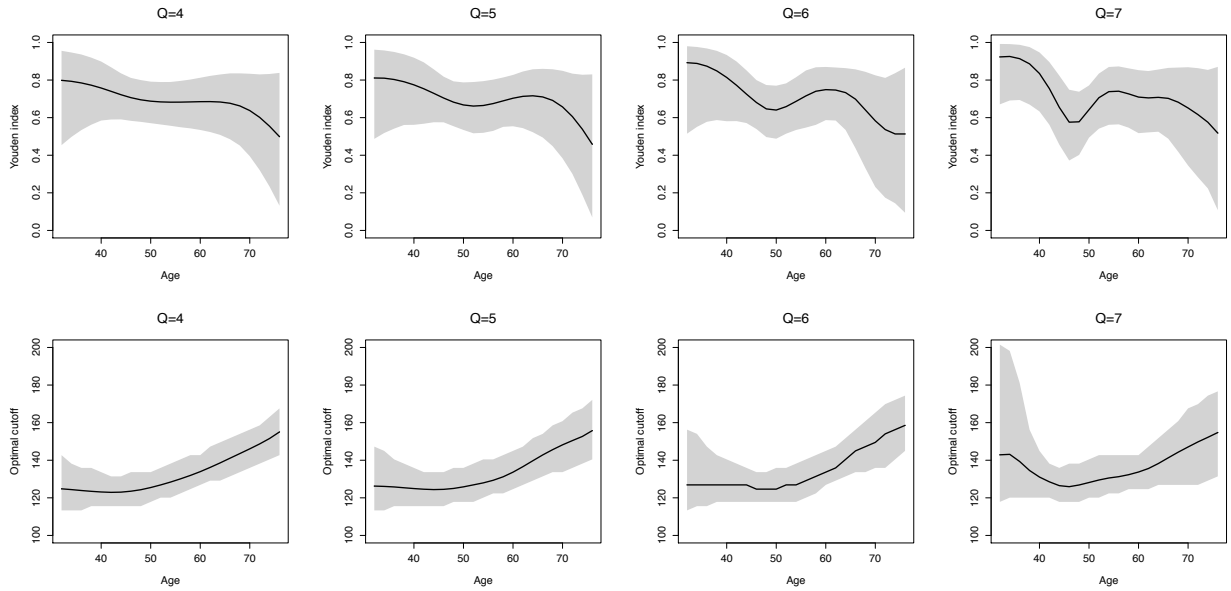


Figure 3: Estimated Youden index and optimal cutoff as a function of age for $Q = 4$, $Q = 5$, $Q = 6$, and $Q = 7$. Solid lines represent posterior means and the gray areas correspond to pointwise 95% posterior bands.

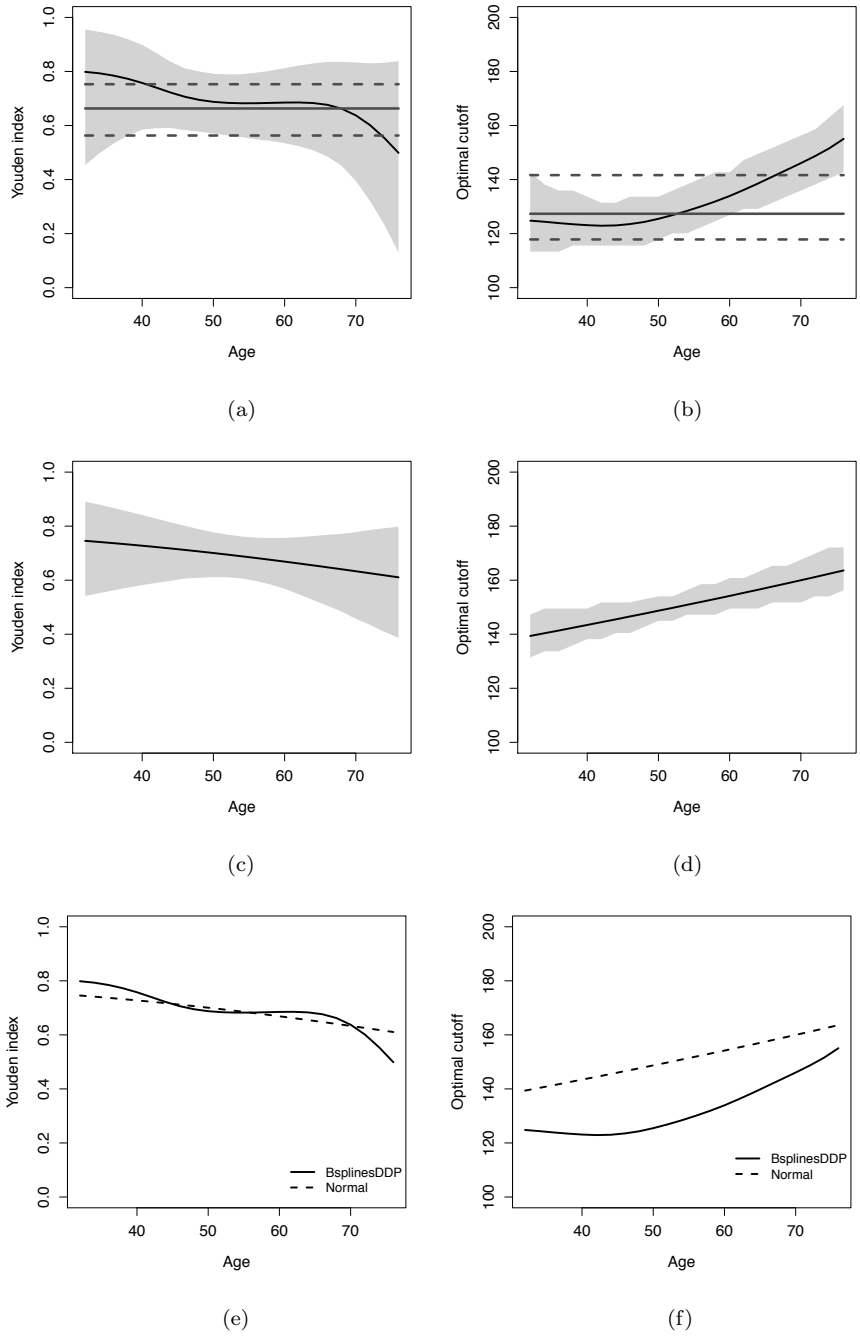


Figure 4: Estimated Youden index and optimal cutoff as a function of age. Panels (a) and (b) present results from the B-splines DDP estimator (along with the results obtained when ignoring the effect of age), while panels (c) and (d) present results obtained under the normal linear model. Solid lines represent posterior means and the gray areas correspond to pointwise 95% posterior bands. For ease of comparison, panels (e) and (f) display the posterior means together.

Biometrics 000, 000–000 DOI: 000
000 0000

Web-based Supplementary Materials for "Nonparametric Bayesian covariate-adjusted estimation of the Youden index" by Vanda Inácio de Carvalho, Miguel de Carvalho and Adam Branscum

This paper has been submitted for consideration for publication in *Biometrics*

In this supplement to the main paper we present computational and technical details, along with supporting figures and other results. Specifically, the full conditional distributions for the Gibbs sampler used to fit the nonparametric regression model are presented in Section A; we fit the unconditional (i.e., without covariates) nonparametric model in Section 2.1 of the paper using analogous computational methods. Section B contains proofs of Theorems 1 and 2. Details on the nonparametric kernel based method of Zhou and Qin (2015) and additional figures are provided in Section C. Finally, in Section D we give details on a data-driven prior and provide additional figures.

Let D and $\bar{D}$ denote the diseased and nondiseased populations, respectively. For $d \in \{D, \bar{D}\}$, the cumulative distribution functions are denoted $F_d(\cdot)$ in the absence of covariates and $F_d(\cdot|\mathbf{x})$ in the regression model with covariates. Following the notation used in the main paper, we write $\text{YI} = \max_{c\in\mathbb{R}}\{F_{\bar{D}}(c) - F_D(c)\}$ to denote the Youden index induced by the Dirichlet Process Mixture (DPM) model in Section 2.1, and $\text{YI}(\mathbf{x}) = \max_{c\in\mathbb{R}}\{F_{\bar{D}}(c \mid \mathbf{x}) - F_D(c \mid \mathbf{x})\}$ to denote the covariate-dependent Youden index induced by the Dependent Dirichlet Process (DDP) model in Section 2.2. Below, we use the notations $\|\cdot\|_\infty$ and $\|\cdot\|_1$ to denote the sup-norm and the L_1-norm, respectively, so that $\|F_{\bar{D}} - F_D\|_\infty = \sup_{c\in\mathbb{R}} |F_{\bar{D}}(c) - F_D(c)|$ and $\|f_{\bar{D}} - f_D\|_1 = \int_{\mathbb{R}} |f_{\bar{D}}(u) - f_D(u)|\mathrm{d}u$.

For the proof of Theorem 1 in Appendix B we use the following auxiliary lemma often know in mathematical language as the *reverse triangle inequality*.

LEMMA 1: *Let V be a normed vector space. Then, $|\,\|\boldsymbol{a}\| - \|\boldsymbol{b}\|\,| \leqslant \|\boldsymbol{a} - \boldsymbol{b}\|$, for all $\boldsymbol{a}$, $\boldsymbol{b}$ in V.*

Proof. See Christensen (2010, Lemma 2.1.2).

Section A: Blocked Gibbs sampler algorithm

The Gibbs sampling algorithm that we used to fit our nonparametric DDP regression model is essentially a covariate-dependent version of the Blocked Gibbs sampler algorithm in Ishwaran and James (2002), where iterative sampling is conducted using the full conditional distributions catalogued below. We omit the subscripts D and $\bar{D}$ in order to present a general setting that is applicable to both the diseased and nondiseased populations. We describe an algorithm for the non-spline version in which the sampling models for the data from both populations have the form $F(\cdot|\mathbf{x}) = \int \Phi(\cdot|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2)\, \mathrm{d}G(\boldsymbol{\beta}, \sigma^2)$, where $G \sim \mathrm{DP}(\alpha, G^*)$ and $G^*(\boldsymbol{\beta}, \sigma^{-2}) = N_p(\boldsymbol{m}_{\boldsymbol{\beta}}, \boldsymbol{S}_{\boldsymbol{\beta}})\Gamma(a, b)$ with $\boldsymbol{m}_{\boldsymbol{\beta}} \sim N_p(\boldsymbol{m}_0, \boldsymbol{S}_0)$ and $\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} \sim \mathrm{Wishart}_p(\nu, (\nu\boldsymbol{\psi})^{-1})$. The truncation

$$G^L(\cdot) = \sum_{\ell=1}^{L} p_\ell \delta_{(\boldsymbol{\beta}_\ell, \sigma_\ell^2)}(\cdot)$$

is used to approximate G (Ishwaran and James, 2001; Ishwaran and Zarepour, 2000, 2002), where L is chosen to be large, e.g., $L = 20$ (Chung and Dunson, 2009). The weights are obtained from the stick-breaking equation

$$p_\ell = q_\ell \prod_{r<\ell}(1 - q_r),$$

where $q_1, \ldots, q_{L-1} \mid \alpha \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$ and $q_L = 1$. Upon introducing membership indicators (Diebolt and Robert, 1994) such that $z_i = \ell$ when y_i comes from $N(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell, \sigma_\ell^2)$, the full conditional distributions are as follows:

$$q_\ell \mid \text{else} \sim \mathrm{Beta}\left(n_\ell + 1, \alpha + \sum_{r=\ell+1}^{L} n_r\right),$$

where $n_\ell = \sum_{i=1}^{n} I(z_i = \ell)$ is the number of observations from component ℓ; in addition, note that $P(z_i = \ell \mid \text{else}) \propto p_\ell \phi(y_i|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell, \sigma_\ell^2)$ and

$$\begin{cases} \boldsymbol{\beta}_\ell \mid \text{else} & \sim N_p\left(\boldsymbol{V}_\ell\left(\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{m}_{\boldsymbol{\beta}} + \sigma_\ell^{-2}\sum_{\{i:z_i=\ell\}}\mathbf{x}_i y_i\right), \boldsymbol{V}_\ell\right), \quad \boldsymbol{V}_\ell = \left(\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} + \sigma_\ell^{-2}\sum_{\{i:z_i=\ell\}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}, \\ \sigma_\ell^{-2} \mid \text{else} & \sim \Gamma\left(a + \frac{n_\ell}{2}, b + \frac{1}{2}\sum_{\{i:z_i=\ell\}}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\ell)^2\right), \end{cases}$$

where

$$\begin{cases} \boldsymbol{m}_{\boldsymbol{\beta}} \mid \text{else} & \sim N_p\left(\boldsymbol{V}\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\sum_{\ell=1}^{L}\boldsymbol{\beta}_\ell\right), \boldsymbol{V}\right), \quad \boldsymbol{V} = (\boldsymbol{S}_0^{-1} + L\boldsymbol{S}_{\boldsymbol{\beta}}^{-1})^{-1}, \\ \boldsymbol{S}_{\boldsymbol{\beta}}^{-1} \mid \text{else} & \sim \text{Wishart}_p\left(\nu + L, \left(\nu\boldsymbol{\psi} + \sum_{\ell=1}^{L}(\boldsymbol{\beta}_\ell - \boldsymbol{m}_{\boldsymbol{\beta}})(\boldsymbol{\beta}_\ell - \boldsymbol{m}_{\boldsymbol{\beta}})^{\mathrm{T}}\right)^{-1}\right), \end{cases}$$

and

$$\alpha \mid \text{else} \sim \text{Gamma}\left(a + L, b - \sum_{\ell=1}^{L-1}\log(1-q_\ell)\right).$$

The full conditional distributions for the spline version follows immediately by replacing $\mathbf{x}^{\mathrm{T}}$ by $\mathbf{z}^{\mathrm{T}} = (B_1(x), \ldots, B_Q(x))$ and by using $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lQ})^{\mathrm{T}}$.

Section B: Proofs of Theorems 1 and 2

Proof of Theorem 1. The proof entails simple manipulations and a result from Lijoi et al. (2004, §3). Note that

$$\begin{aligned} |\mathrm{YI}_\omega - \mathrm{YI}| &= \left|\max_{c\in\mathbb{R}}\{F_{\bar{D}}^{\omega}(c) - F_D^{\omega}(c)\} - \max_{c\in\mathbb{R}}\{F_{\bar{D}}(c) - F_D(c)\}\right| \\ &= |\,\|F_{\bar{D}}^{\omega} - F_D^{\omega}\|_\infty - \|F_{\bar{D}} - F_D\|_\infty\,| \\ &\leqslant \|(F_{\bar{D}}^{\omega} - F_D^{\omega}) - (F_{\bar{D}} - F_D)\|_\infty \\ &= \|(F_{\bar{D}}^{\omega} - F_{\bar{D}}) - (F_D^{\omega} - F_D)\|_\infty \end{aligned} \tag{1}$$

where the penultimate step follows from the reverse triangle inequality. It follows from (1) that

$$|\mathrm{YI}_\omega - \mathrm{YI}| \leqslant \|F_{\bar{D}}^{\omega} - F_{\bar{D}}\|_\infty + \|F_D^{\omega} - F_D\|_\infty \leqslant \|f_{\bar{D}}^{\omega} - f_{\bar{D}}\|_1 + \|f_D^{\omega} - f_D\|_1, \tag{2}$$

where the last step follows from

$$\begin{aligned} \|F_{\bar{D}}^{\omega} - F_{\bar{D}}\|_\infty = \sup_{c\in\mathbb{R}}\left|\int_{-\infty}^{c}\{f_{\bar{D}}^{\omega}(u) - f_{\bar{D}}(u)\}\,\mathrm{d}u\right| &\leqslant \sup_{c\in\mathbb{R}}\int_{-\infty}^{c}|f_{\bar{D}}^{\omega}(u) - f_{\bar{D}}(u)|\,\mathrm{d}u \\ &= \int_{-\infty}^{\infty}|f_{\bar{D}}^{\omega}(u) - f_{\bar{D}}(u)|\,\mathrm{d}u = \|f_{\bar{D}}^{\omega} - f_{\bar{D}}\|_1, \end{aligned}$$

and analogously $\|F_D^{\omega} - F_D\|_\infty \leqslant \|f_D^{\omega} - f_D\|_1$.

Hence, as it can be noticed from (2), to have $|\mathrm{YI}_\omega - \mathrm{YI}| < \varepsilon$, it would suffice having $\|f_{\bar{D}}^{\omega} - f_{\bar{D}}\|_1 < \varepsilon/2$ and $\|f_D^{\omega} - f_D\|_1 < \varepsilon/2$, thus implying that

$$\{\omega\in\Omega : |\mathrm{YI}_\omega - \mathrm{YI}| < \varepsilon\} \supseteq \{\omega\in\Omega : \|f_{\bar{D}}^{\omega} - f_{\bar{D}}\|_1 < \varepsilon/2, \|f_D^{\omega} - f_D\|_1 < \varepsilon/2\},$$

from where it finally follows that

$$P(\omega \in \Omega : |\mathrm{YI}_\omega - \mathrm{YI}| < \varepsilon) \geqslant P(\omega \in \Omega : \|f_D^\omega - f_{\bar{D}}\|_1 < \varepsilon/2) \times P(\omega \in \Omega : \|f_D^\omega - f_D\|_1 < \varepsilon/2) > 0,$$

for every $\varepsilon > 0$, given that as a consequence of Lijoi et al. (2004, §3), it holds that for every $\varepsilon > 0$,

$$P(\omega \in \Omega : \|f_{\bar{D}}^\omega - f_{\bar{D}}\|_1 < \varepsilon/2) > 0, \quad P(\omega \in \Omega : \|f_D^\omega - f_D\|_1 < \varepsilon/2) > 0.$$

■

Proof of Theorem 2. The proof is analogous to that of Theorem 1, but we need to use Theorem 4 in Barrientos et al. (2012), which is essentially a predictor-dependent version of the results in Lijoi et al. (2004, §3). Similar arguments as in the proof of Theorem 1 yield,

$$|\mathrm{YI}_\omega(\boldsymbol{x}_i) - \mathrm{YI}(\boldsymbol{x}_i)| \leqslant \|f_{\bar{D}}^\omega(\cdot \mid \boldsymbol{x}_i) - f_{\bar{D}}(\cdot \mid \boldsymbol{x}_i)\|_1 + \|f_D^\omega(\cdot \mid \boldsymbol{x}_i) - f_D(\cdot \mid \boldsymbol{x}_i)\|_1, \quad i = 1, \ldots, n.$$

Hence, to have $|\mathrm{YI}_\omega(\boldsymbol{x}_i) - \mathrm{YI}(\boldsymbol{x}_i)| < \varepsilon$, it would suffice having

$$\|f_{\bar{D}}^\omega(\cdot \mid \boldsymbol{x}_i) - f_{\bar{D}}(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2, \quad \|f_D^\omega(\cdot \mid \boldsymbol{x}_i) - f_D(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2, \quad i = 1, \ldots, n.$$

By similar arguments as in the proof of Theorem 1,

$$\begin{aligned} &P(\omega \in \Omega : |\mathrm{YI}_\omega(\boldsymbol{x}_i) - \mathrm{YI}(\boldsymbol{x}_i)| < \varepsilon,\, i = 1, \ldots, n) \\ &\geqslant P(\omega \in \Omega : \|f_{\bar{D}}^\omega(\cdot \mid \boldsymbol{x}_i) - f_{\bar{D}}(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2,\, i = 1, \ldots, n) \\ &\quad \times P(\omega \in \Omega : \|f_D^\omega(\cdot \mid \boldsymbol{x}_i) - f_D(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2,\, i = 1, \ldots, n) > 0, \end{aligned}$$

for every $\varepsilon > 0$, given that as a consequence of results on the Hellinger support of the DDP (Barrientos et al., 2012, Theorem 4), it holds that for every $\varepsilon > 0$,

$$\begin{cases} P(\omega \in \Omega : \|f_{\bar{D}}^\omega(\cdot \mid \boldsymbol{x}_i) - f_{\bar{D}}(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2,\, i = 1, \ldots, n) > 0, \\ P(\omega \in \Omega : \|f_D^\omega(\cdot \mid \boldsymbol{x}_i) - f_D(\cdot \mid \boldsymbol{x}_i)\|_1 < \varepsilon/2,\, i = 1, \ldots, n) > 0. \end{cases}$$

■

Section C: Simulation study details and figures

In this section we present additional details and supporting figures to the statistical analysis conducted in the simulation study section of the main paper.

Nonparametric kernel method (Zhou and Qin, 2015). This method is based on modeling test outcomes through nonparametric heteroscedastic regression models

$$y_D = \mu_D(x) + \sigma_D(x)\varepsilon_D, \qquad y_{\bar{D}} = \mu_{\bar{D}}(x) + \sigma_{\bar{D}}(x)\varepsilon_{\bar{D}},$$

where μ_D and $\mu_{\bar{D}}$ are the regression functions and σ_D^2 and $\sigma_{\bar{D}}^2$ are the variance functions. Here ε_D and $\varepsilon_{\bar{D}}$ are independent random variables with zero mean, variance one, and distribution functions F_{ε_D} and $F_{\varepsilon_{\bar{D}}}$, respectively. Both the regression and variance functions in each group are estimated using local constant estimators; that is, using Nadaraya–Watson estimators (Fan and Gijbels, 1996, Section 2). The bandwidths involved in the computation of these estimators were selected sequentially and by expected Kullback–Leibler cross-validation (Hurvich et al., 1998) as implemented in the `R` function `npregbw` from the `np` package. Once we have the estimates $\widehat{\mu}_d(x)$ and $\widehat{\sigma}_d^2(x)$, $d \in \{D, \bar{D}\}$, we can estimate the standardized residuals

$$\widehat{\varepsilon}_{Di} = \frac{y_{Di} - \widehat{\mu}_D(x_{Di})}{\widehat{\sigma}_D(x_{Di})}, \qquad \widehat{\varepsilon}_{\bar{D}j} = \frac{y_{\bar{D}j} - \widehat{\mu}_{\bar{D}}(x_{\bar{D}j})}{\widehat{\sigma}_{\bar{D}}(x_{\bar{D}j})}, \quad i = 1, \ldots, n_D, \quad j = 1, \ldots, n_{\bar{D}}.$$

Denoting by $\widehat{F}_{\widehat{\varepsilon}_D}$ and $\widehat{F}_{\widehat{\varepsilon}_{\bar{D}}}$, the empirical distribution functions of $\hat{\varepsilon}_D$ and $\widehat{\varepsilon}_{\bar{D}}$, respectively, we have

$$\widehat{\mathrm{YI}}_{\mathrm{kernel}}(x) = \max_{c \in \mathbb{R}} \left\{ \widehat{F}_{\widehat{\varepsilon}_{\bar{D}}}\left(\frac{c - \widehat{\mu}_{\bar{D}}(x)}{\widehat{\sigma}_{\bar{D}}(x)} \right) - \widehat{F}_{\widehat{\varepsilon}_D}\left(\frac{c - \widehat{\mu}_D(x)}{\widehat{\sigma}_D(x)} \right) \right\},$$
$$\hat{c}^*_{\mathrm{kernel}}(x) = \mathrm{argmax}_{c \in \mathbb{R}} \left\{ \widehat{F}_{\widehat{\varepsilon}_{\bar{D}}}\left(\frac{c - \widehat{\mu}_{\bar{D}}(x)}{\widehat{\sigma}_{\bar{D}}(x)} \right) - \widehat{F}_{\widehat{\varepsilon}_D}\left(\frac{c - \widehat{\mu}_D(x)}{\widehat{\sigma}_D(x)} \right) \right\}.$$

To obtain pointwise confidence bands, a bootstrap of the residuals (see, for instance, Gonzalez-Manteiga et al., 2011, Section 5) is employed.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Table 1 about here.]

Section D: Application—Data-driven prior and supplementary figures

A sensitivity analysis was conducted using the following prior specification suggested by Inácio de Carvalho et al. (2013) and Zhou et al. (2015)

$$a_d = 3, \quad b_d = \widehat{\sigma}_d^2,$$

$$\mathbf{m}_{d0} = (\mathbf{z}_d'\mathbf{z}_d)^{-1}\mathbf{z}_d'\mathbf{y}_d, \quad \mathbf{S}_{d0} = \widehat{\sigma}_d^2(\mathbf{z}_d'\mathbf{z}_d)^{-1},$$

$$\nu_d = Q + 2, \quad \Psi_d = 30\mathbf{S}_{d0},$$

with $\widehat{\sigma}_d^2 = \|\mathbf{y}_d - \mathbf{z}_d\mathbf{m}_{d0}\|^2/(n_d - Q - 1)$. The results are presented in Figures 15 and 16 and in Table 2.

[Figure 11 about here.]

[Figure 12 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Table 2 about here.]

Section E: `R` code for implementing our methods

Below, we discuss some `R` code for illustrating how to implement the nonparametric Bayes estimator for the covariate-adjusted youden index and corresponding optimal cutoff. Before running the code chunks below, start by cleaning workspace and install the following packages (if not installed).

```
rm(list = ls())
if (!require("splines")) install.packages("splines")
if (!require("Hmisc")) install.packages("Hmisc")
if (!require("MASS")) install.packages("MASS")
```

For reproducibility reasons, this pdf file has been prepared using `knitr` (Xie, 2015); we fix `setseed` and list below the information about `R`, the OS, and loaded packages:

```
set.seed(1)
sessionInfo()
```

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: macOS Sierra 10.12.1
##
## locale:
## [1] en_US.UTF-8/C/en_US.UTF-8/C/C/en_US.UTF-8
##
## attached base packages:
## [1] splines   stats     graphics  grDevices utils     datasets  base
##
## other attached packages:
## [1] MASS_7.3-45     Hmisc_3.17-4    ggplot2_2.1.0   Formula_1.2-1
## [5] survival_2.39-5 lattice_0.20-34 knitr_1.15
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5         cluster_2.0.5       magrittr_1.5
##  [4] munsell_0.4.3       colorspace_1.2-6    highr_0.6
##  [7] stringr_1.0.0       plyr_1.8.4          tools_3.3.2
```

```
## [10] nnet_7.3-12          grid_3.3.2           data.table_1.9.6
## [13] gtable_0.2.0         latticeExtra_0.6-28  Matrix_1.2-7.1
## [16] gridExtra_2.2.1      RColorBrewer_1.1-2   acepack_1.3-3.3
## [19] rpart_4.1-10         evaluate_0.10        stringi_1.1.1
## [22] methods_3.3.2        scales_0.4.0         chron_2.3-47
## [25] foreign_0.8-67
```

In the code chunks below, we follow the 80 characters per line standard. The key function for fitting the covariate-adjusted Youden index is:

```
bsplinesddp <- function(y, x, grid, xpred, m, S, nu, psi, atau, btau,
                        alpha, L, nsim, knots) {
  yt <- y / sd(y)
  n <- length(y)
  ngrid <- length(grid)
  npred <- length(xpred)
  X <- bs(x, degree = 3, knots = knots, intercept = TRUE)
  k <- ncol(X)
  Xpred <- predict(bs(x, degree = 3, knots = knots, intercept = TRUE),
                   xpred)

  p <- ns <- rep(0, L)
  v <- rep(1 / L, L)
  v[L] <- 1

  beta <- matrix(0, nrow = L, ncol = k)
  tau <- rep(1 / var(yt), L)

  prop <- prob <- matrix(0, nrow = n, ncol = L)

  P <- Tau <- Sigma2 <- matrix(0, nrow = nsim, ncol = L)
  Beta <- Beta1 <- array(0, c(nsim, L, k))
  Beta[1, , ] <- beta
  Tau[1, ] <- tau

  mu <- matrix(0, nrow = nsim, ncol = k)
  Sigmainv <- array(0, c(nsim, k, k))
  mu[1, ] <- mvrnorm(1, mu = m, Sigma = S)
  Sigmainv[1, , ] <- rWishart(1, df = nu, solve(nu * psi))

  Dens <- array(0, c(nsim, ngrid, L, npred))
  Densm <- array(0, c(nsim, ngrid, npred))
  Fdist <- array(0, c(nsim, ngrid, L, npred))
```

```
Fdistm <- array(0, c(nsim, ngrid, npred))

## 1) ALLOCATE EACH OBSERVATION TO A COMPONENT MIXTURE
for(i in 2:nsim) {
  cumv <- cumprod(1 - v)

  p[1] <- v[1]
  for(l in 2:L)
    p[l] <- v[l] * cumv[l - 1]
  for(l in 1:L)
    prop[, l] <- p[l] * dnorm(yt, mean = X %*% beta[l, ],
                              sd = sqrt(1 / tau[l]))
  prob <- prop / apply(prop, 1, sum)
  z <- rMultinom(prob, 1)
  P[i, ] <- p

  for(l in 1:L)
    ns[l] <- length(which(z == l))

  ## 2) UPDATE STICK-BREAKING WEIGHTS
  for(l in 1:(L - 1))
    v[l] <- rbeta(1, 1 + ns[l], alpha + sum(ns[(l + 1):L]))

  ## 3) UPDATE PARAMETERS OF EACH COMPONENT MIXTURE
  for(l in 1:L) {
    tX <- matrix(t(X[z == l, ]), nrow = k, ncol = ns[l])
    V <- solve(Sigmainv[i - 1, ,] + tau[l] * tX %*% X[z == l, ])
    mu1 <- V %*% (Sigmainv[i - 1, , ] %*% mu[i - 1, ] + tau[l] *
                  tX %*% yt[z == l])
    Beta[i, l, ] <- beta[l,] <- mvrnorm(1, mu = mu1, Sigma = V)
    Beta1[i, l, ] <- sd(y) * Beta[i, l, ]

    Tau[i, l] <- tau[l] <- rgamma(1, shape = atau + (ns[l] / 2),
                                  rate = btau + 0.5 * (t(yt[z == l] -
                                  X[z == l, ] %*% beta[l, ]) %*%
                                  (yt[z == l] - X[z == l, ]
                                  %*% beta[l, ])))
    Sigma2[i, l] <- var(y) * (1 / Tau[i, l])
  }

  Vaux <- solve(solve(S) + L * Sigmainv[i - 1, , ])
  meanmu <- Vaux %*% (solve(S) %*% m + Sigmainv[i - 1, , ] %*%
                    t(t(apply(Beta[i, , ], 2, sum))))
  mu[i, ] <- mvrnorm(1, mu = meanmu, Sigma = Vaux)
```

```
    Vaux1 <- 0
    for(l in 1:L)
      Vaux1 <- Vaux1 + (Beta[i, l, ] - mu[i, ]) %*%
               t((Beta[i, l, ] - mu[i, ]))
    Sigmainv[i, , ] <- rWishart(1, nu + L, solve(nu * psi + Vaux1))

    ## 4) COMPUTE DENSITY AND DISTRIBUTION FUNCTION TRAJECTORIES
    for(l in 1:L) {
      for(j in 1:npred) {
        Dens[i, ,l , j] <- P[i, l] * dnorm(grid, Xpred[j, ] %*%
                                           Beta1[i, l, ],
                                           sqrt(Sigma2[i, l]))
        Fdist[i, , l, j] <- P[i, l] * pnorm(grid, Xpred[j, ] %*%
                                              Beta1[i, l, ],
                                              sqrt(Sigma2[i, l]))
      }
    }

    for(j in 1:ngrid) {
      for(l in 1:npred) {
        Densm[i, j, l] <- sum(Dens[i, j, , l])
        Fdistm[i, j, l] <- sum(Fdist[i, j, , l])
      }
    }
  }
  return(list(P, Beta1, Sigma2, Densm, Fdistm))
}
```

The arguments of the `bsplinesddp` function are self-explanatory from the article. To analyze the diabetes data we proceed as follows:

```
## setwd("Add your working directory here")
load("diabetes.Rdata")
ind0 <- which(diabetes[, 2] == 0)
ind1 <- which(diabetes[, 2] == 1)
n0 <- length(ind0)
n1 <- length(ind1)
y0 <- diabetes[ind0, 1]
y1 <- diabetes[ind1, 1]
x0 <- diabetes[ind0, 3]
x1 <- diabetes[ind1, 3]
var0 <- var1 <- 1
```

```
knots0 <- c()
knots1 <- c()
nk <- length(knots0)
```

Next, we apply the workhorse function `bsplinesddp` to the pair (glucose levels, age) for diseased (`x1`, `y1`) and nondiseased subjects (`x0`, `y0`):

```
res0np <- bsplinesddp(y = y0, x = x0, grid = seq(50, 500, len = 200),
                      xpred = seq(32, 76, by = 2), m = rep(0, nk + 4),
                      S = 100 * diag(nk + 4), nu = nk + 6,
                      psi = diag(nk + 4), atau = 0.1, btau = 0.1, alpha = 1,
                      L = 20, nsim = 5000, knots = knots0)
res1np <- bsplinesddp(y = y1, x = x1, grid = seq(50, 500, len = 200),
                      xpred = seq(32, 76, by = 2), m = rep(0, nk + 4),
                      S = 100 * diag(nk + 4), nu = nk + 6,
                      psi = diag(nk + 4), atau = 0.1, btau = 0.1, alpha = 1,
                      L = 20, nsim = 5000, knots = knots1)
```

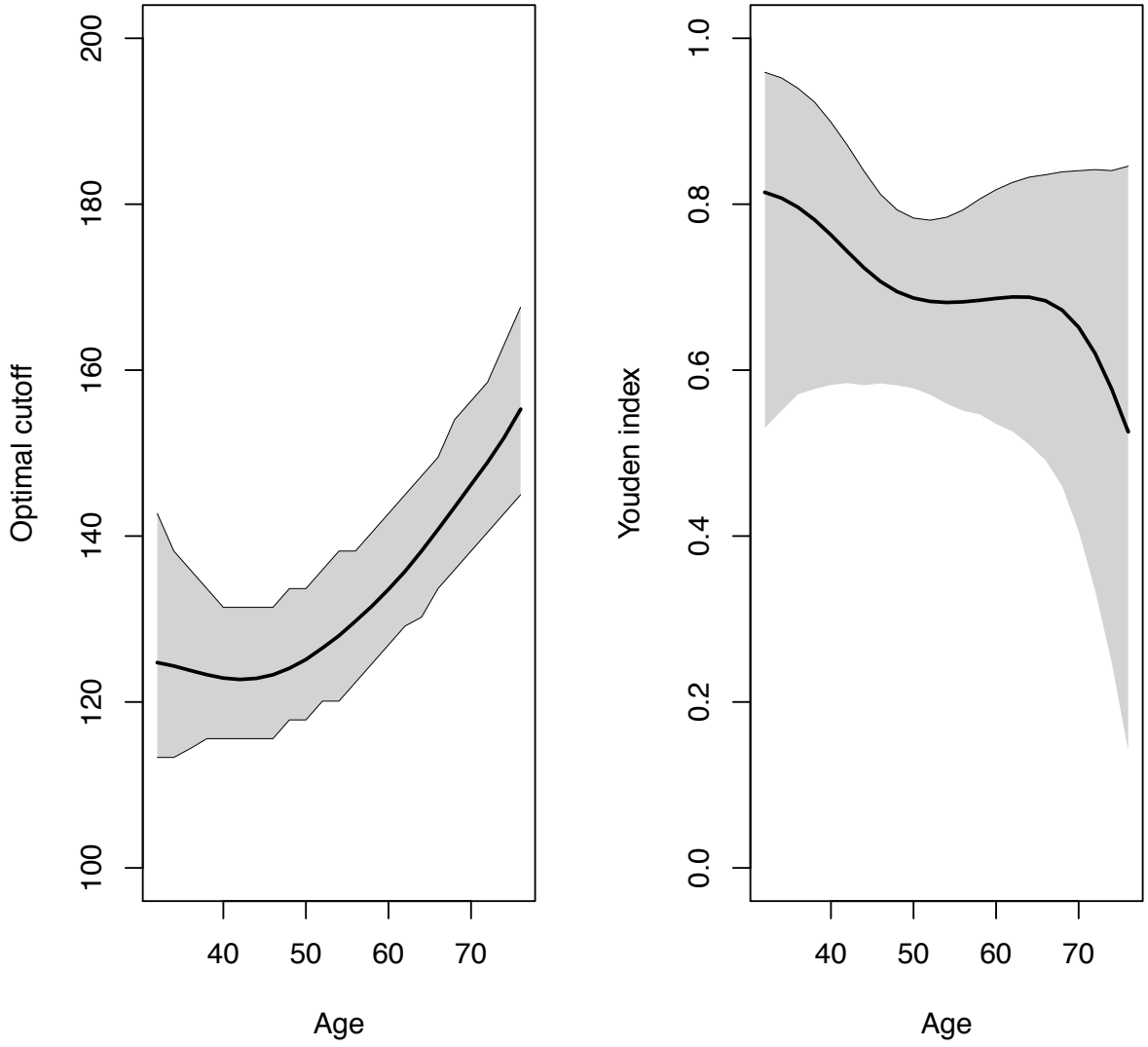
We compute the covariate-adjusted optimal cutoff and corresponding Youden index as follows:

```
grid <- seq(50, 500, len = 200)
xpred <- seq(32, 76, by = 2)
ngrid <- length(grid)
npred <- length(xpred)
nsim <- 5000
nburn <- 1500

difcnp <- array(0, c(nsim - nburn, ngrid, npred))
for(k in 1:npred)
  for(j in 1:ngrid)
  difcnp[, j, k] <- res0np[[5]][(nburn + 1):nsim, j, k] -
                    res1np[[5]][(nburn + 1):nsim, j, k]

coptcnp <- matrix(0, nrow = nsim - nburn, ncol = npred)
for(k in 1:npred)
  for(j in 1:(nsim - nburn))
    coptcnp[j, k] = mean(grid[which(difcnp[j, , k] == max(difcnp[j, , k]))])

coptcrnp <- matrix(nrow = npred, ncol = 3)
for(j in 1:npred) {
```

```
  coptcrnp[j, 1] <- quantile(coptcnp[, j], 0.025)
  coptcrnp[j, 2] <- mean(coptcnp[, j])
  coptcrnp[j, 3] <- quantile(coptcnp[, j], 0.975)
}

yicnp <- matrix(0, nrow = nsim - nburn, ncol = npred)
for(k in 1:npred)
  for(j in 1:(nsim - nburn))
    yicnp[j,k] <- max(difcnp[j, , k])

yicrnp <- matrix(nrow = npred,ncol = 3)
for(j in 1:npred) {
  yicrnp[j, 1] <- quantile(yicnp[, j], 0.025)
  yicrnp[j, 2] <- mean(yicnp[, j])
  yicrnp[j, 3] <- quantile(yicnp[, j], 0.975)
}
```

The first column of Figure 3 in the paper can then be produced using the following lines of code.

```
par(mfrow = c(1, 2))
plot(xpred, coptcrnp[, 2], type = "l", xlim = c(32, 76), ylim = c(100, 200),
     xlab = "Age", ylab = "Optimal cutoff", lwd = 2)
polygon(x = c(rev(xpred), xpred), y = c(rev(coptcrnp[, 1]), coptcrnp[, 3]),
        border = NA, col = "lightgray")
lines(xpred, coptcrnp[, 1], lwd = 0.3)
lines(xpred, coptcrnp[, 2], lwd = 2)
lines(xpred, coptcrnp[, 3], lwd = 0.3)

plot(xpred, yicrnp[, 2], type = "l", xlim = c(32, 76), ylim = c(0, 1),
     xlab = "Age", ylab = "Youden index", lwd = 2)
polygon(x = c(rev(xpred), xpred), y = c(rev(yicrnp[, 1]), yicrnp[, 3]),
        border = NA, col = "lightgray")
lines(xpred, coptcrnp[, 1], lwd = 0.3)
lines(xpred, yicrnp[, 2], lwd = 2)
lines(xpred, yicrnp[, 3], lwd = 0.3)
```



The Log PseudoMarginal Likelihood (LPML) was computed as follows:

```
L <- 20
X0 <- bs(x0, degree = 3, knots = knots0, intercept = TRUE)

term0 <- array(0, c(nsim - nburn, L, n0))
for(i in 1:n0)
  for(k in 1:(nsim-nburn))
    for(l in 1:L)
      term0[k, l, i] <- res0np[[1]][k + nburn,l] *
                        dnorm(y0[i], mean = X0[i,] %*%
```

```
                                res0np[[2]][k + nburn, 1,],
                                sd = sqrt(res0np[[3]][k + nburn, 1]))

termsum0 <- matrix(0, nrow = nsim - nburn, ncol = n0)
for(i in 1:n0)
  for(k in 1:(nsim-nburn))
    termsum0[k, i] <- sum(term0[k, , i])

cpoinv0 <- numeric(n0)
for(i in 1:n0)
  cpoinv0[i] <- mean(1 / termsum0[, i])

cpo0 <- 1 / cpoinv0
lpml0 <- sum(log(cpo0))
```

And we thus get that the value of `lpml0` is −878.18.

References

Barrientos, A. F., Jara, A., and Quintana, F. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis* **7**, 277–310.

Christensen, O. (2010). *Functions, Spaces, and Expansions: Mathematical Tools in Physics and Engineering*. Berlin: Birkhäuser.

Chung, Y., and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* **104**, 1646–1660.

Diebolt, J., and Robert, C. P. (1994). Estimate of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society*, Ser. B **56**, 363–375.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall, London.

Gonzalez-Manteiga, W., Pardo-Fernandez, J. C., and Van Keilegom, I. (2011). ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics* **38** 169–184.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society*, Ser. B **60**, 271–293.

Inácio de Carvalho, V., Jara, A., Hanson, T. E., and de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling. *Bayesian Analysis* **8**, 623–646.

Ishwaran, I., and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96,** 161–173.

Ishwaran, H., and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.

Ishwaran, I., and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. *Biometrika* **87**, 371–390.

Ishwaran, I., and Zarepour, M. (2002). Exact and approximate sum representation for the Dirichlet process. *Canadian Journal of Statistics*, **30**, 269–283.

Lijoi, A., Prünster, I., and Walker, S. G. (2004). Extending Doob's consistency theorem to nonparametric densities. *Bernoulli* **10**, 651–663.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Boca Raton, FL: Chapman & Hall/CRC.

Zhou, H., and Qin, G. (2015). Nonparametric covariate adjustment for the Youden index. In *Applied Statistics in Biomedicine and Clinical Trials Design*. Cham: Springer.

Zhou, A., Hanson, T., and Knapp, R. (2015). Marginal Bayesian nonparametric model for time to disease arrival of threatened amphibian populations. *Biometrics* **71**, 1101–1110.

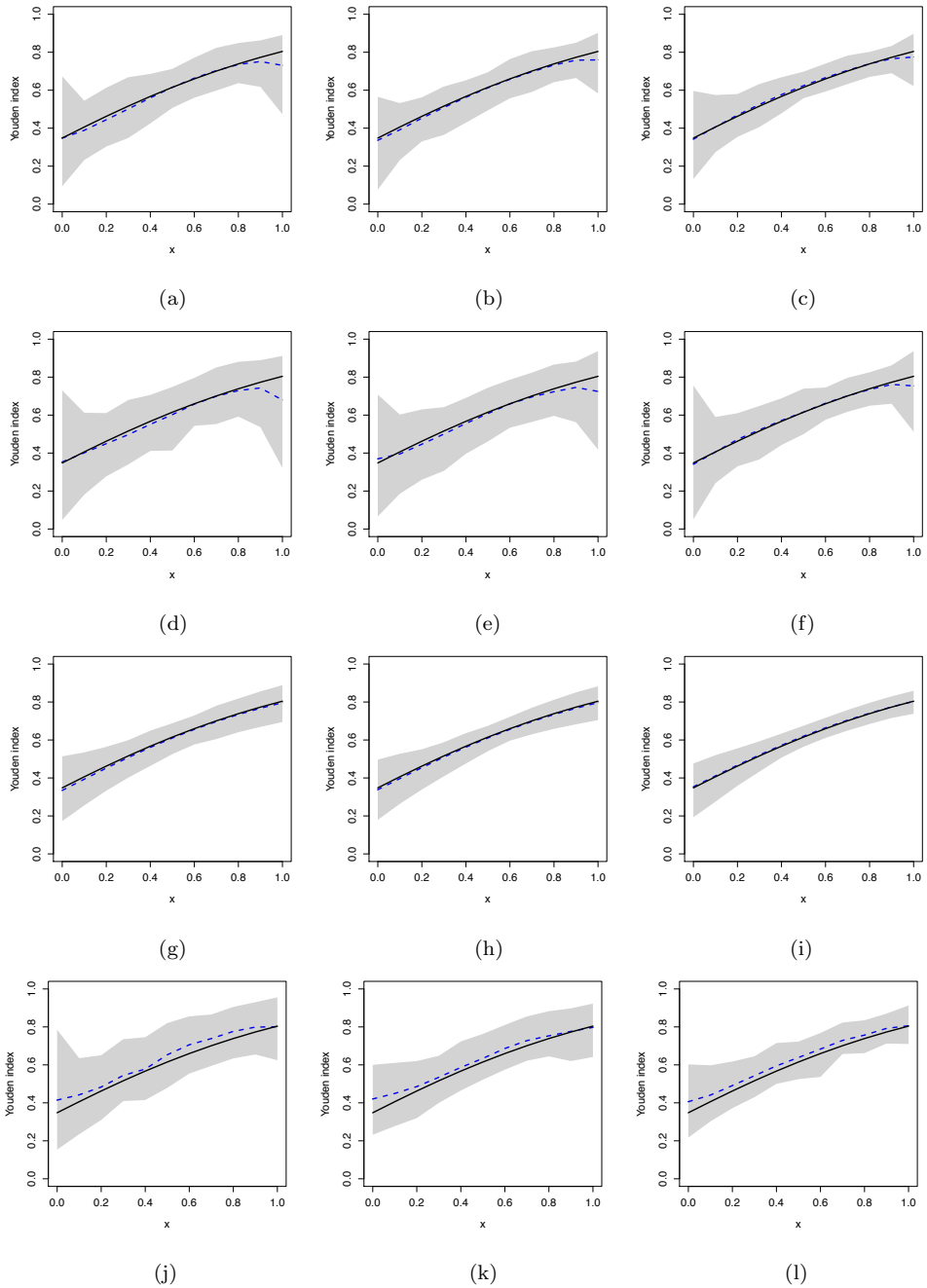


Figure 1. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the Youden index function under Scenario 1. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

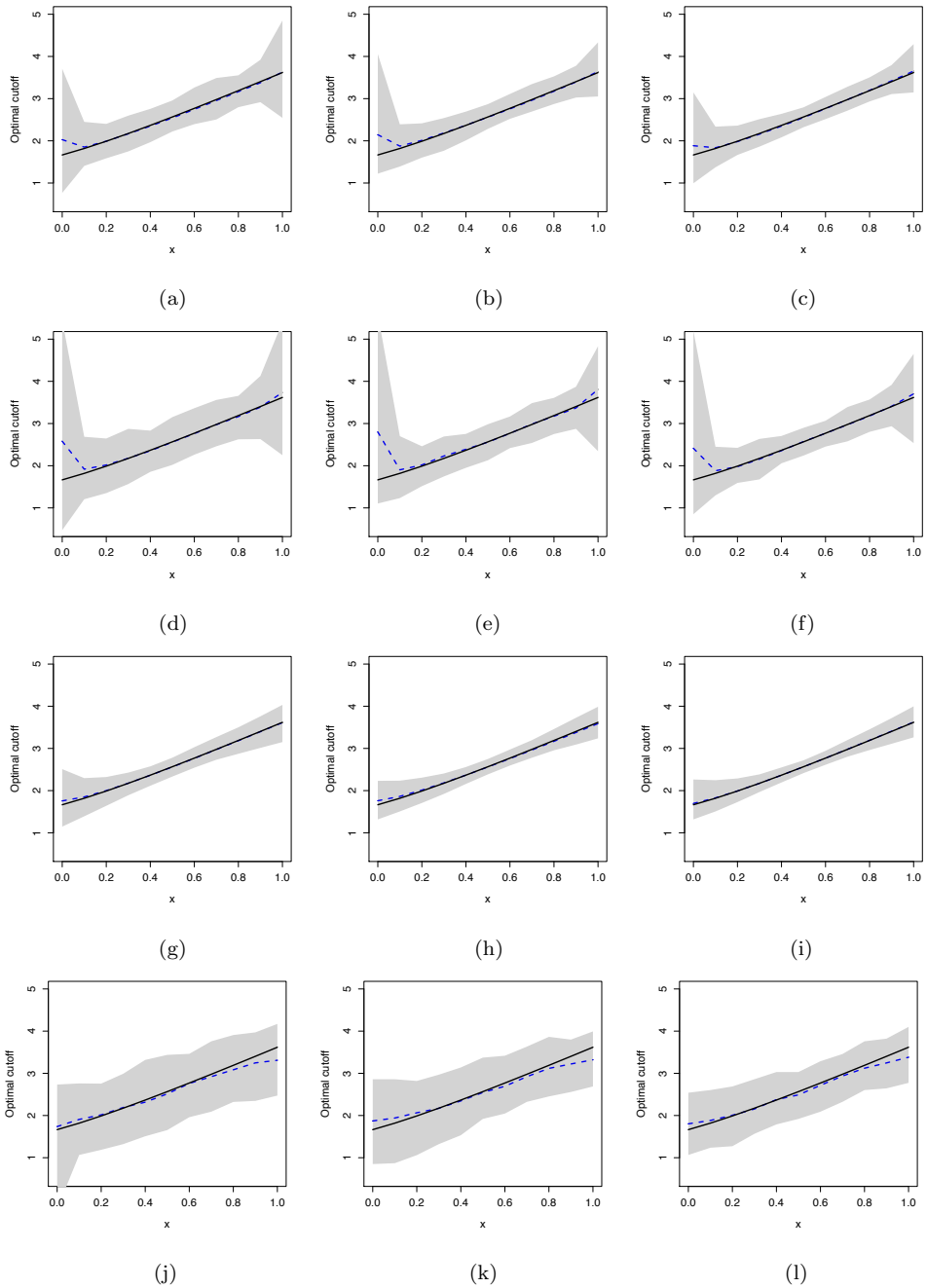


Figure 2. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function under Scenario 1. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

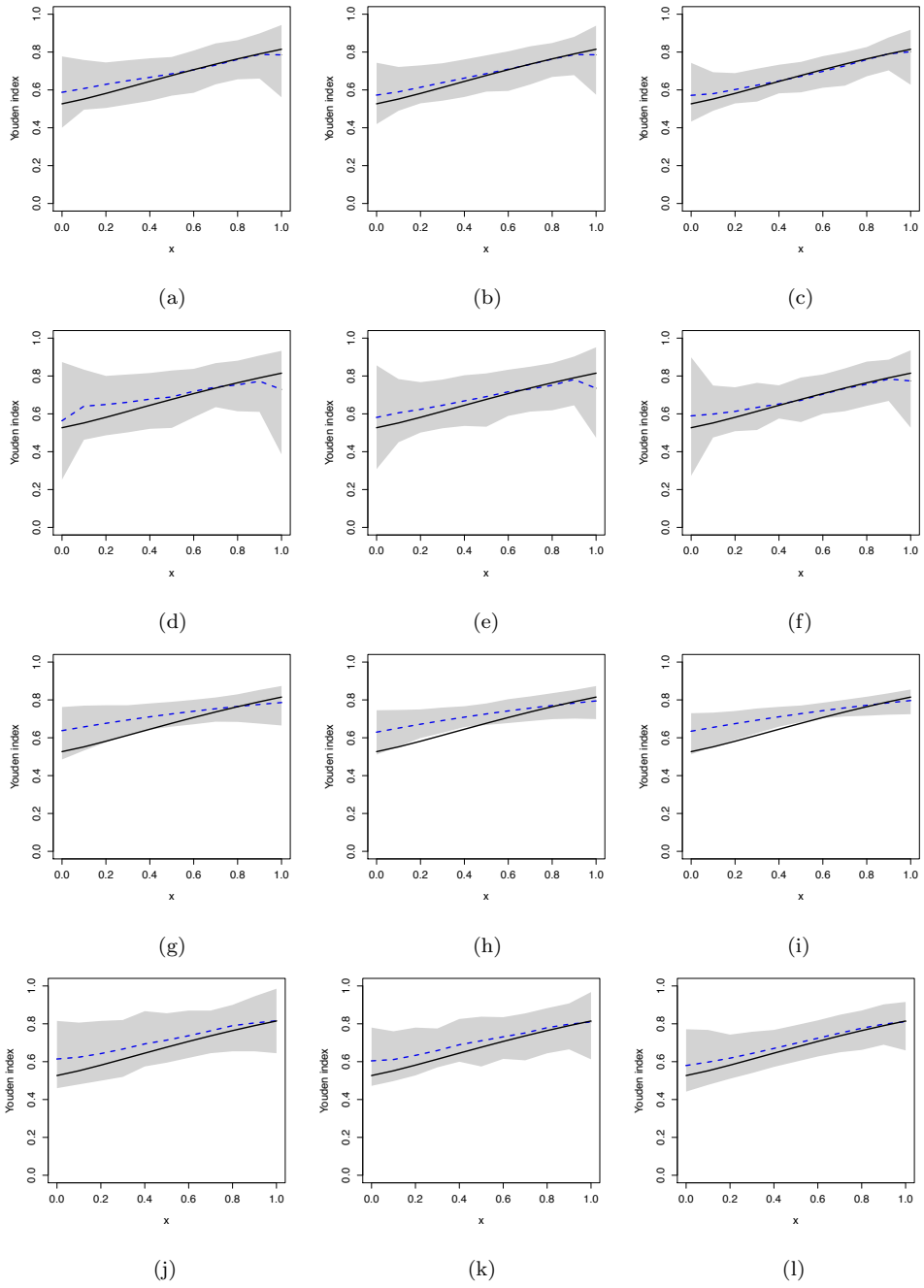


Figure 3. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the Youden index function under Scenario 2. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

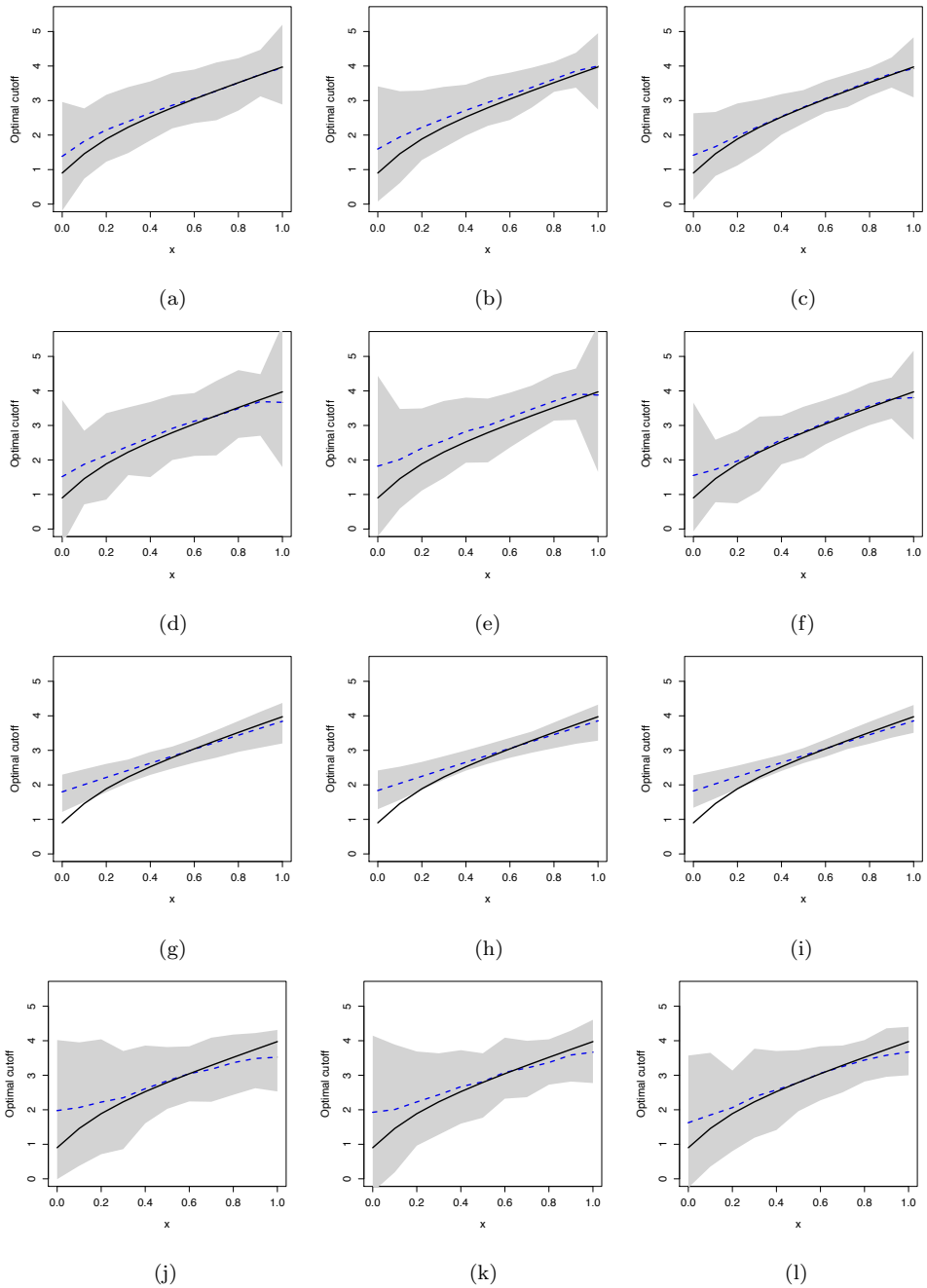


Figure 4. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function under Scenario 2. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

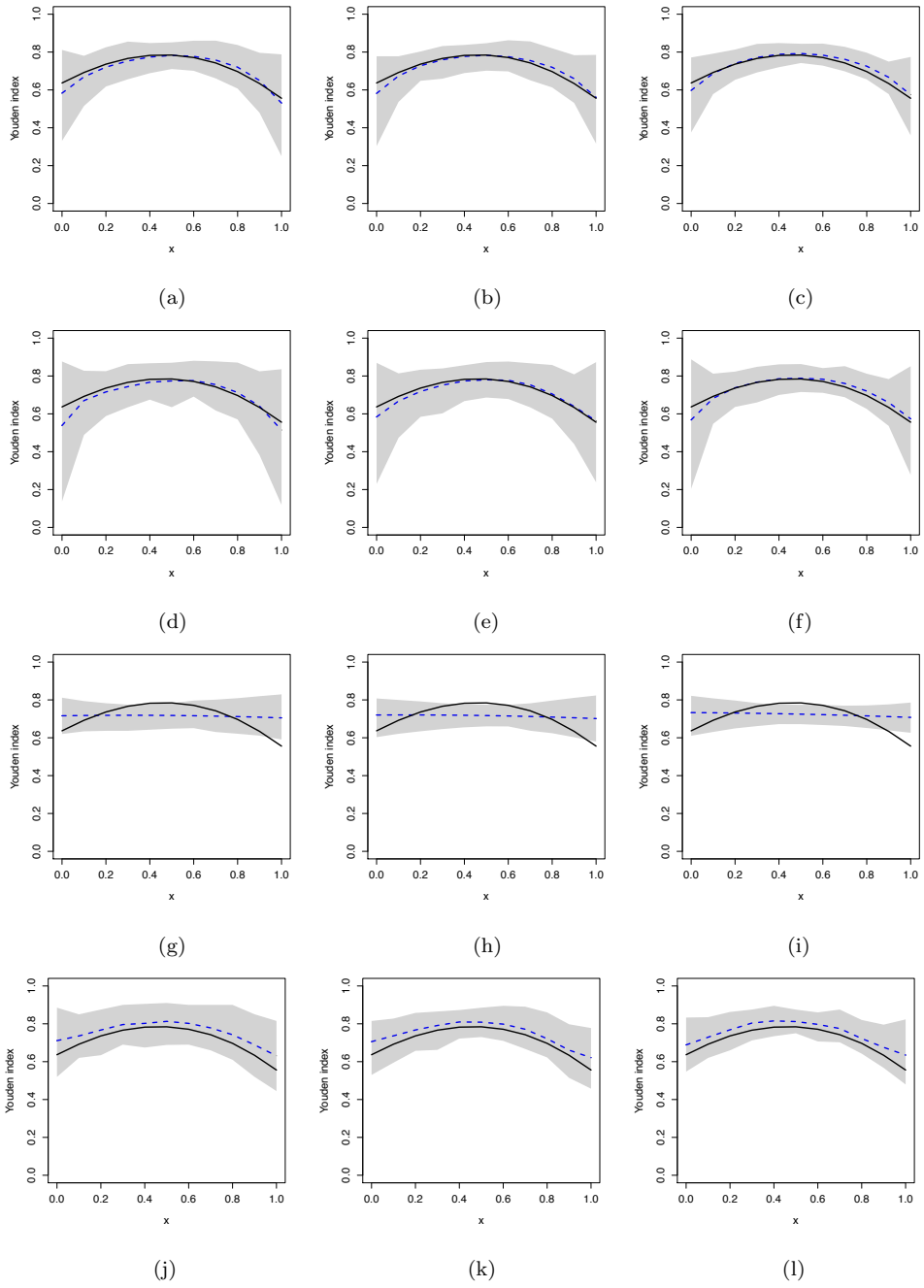


Figure 5. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the Youden index function under Scenario 3. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

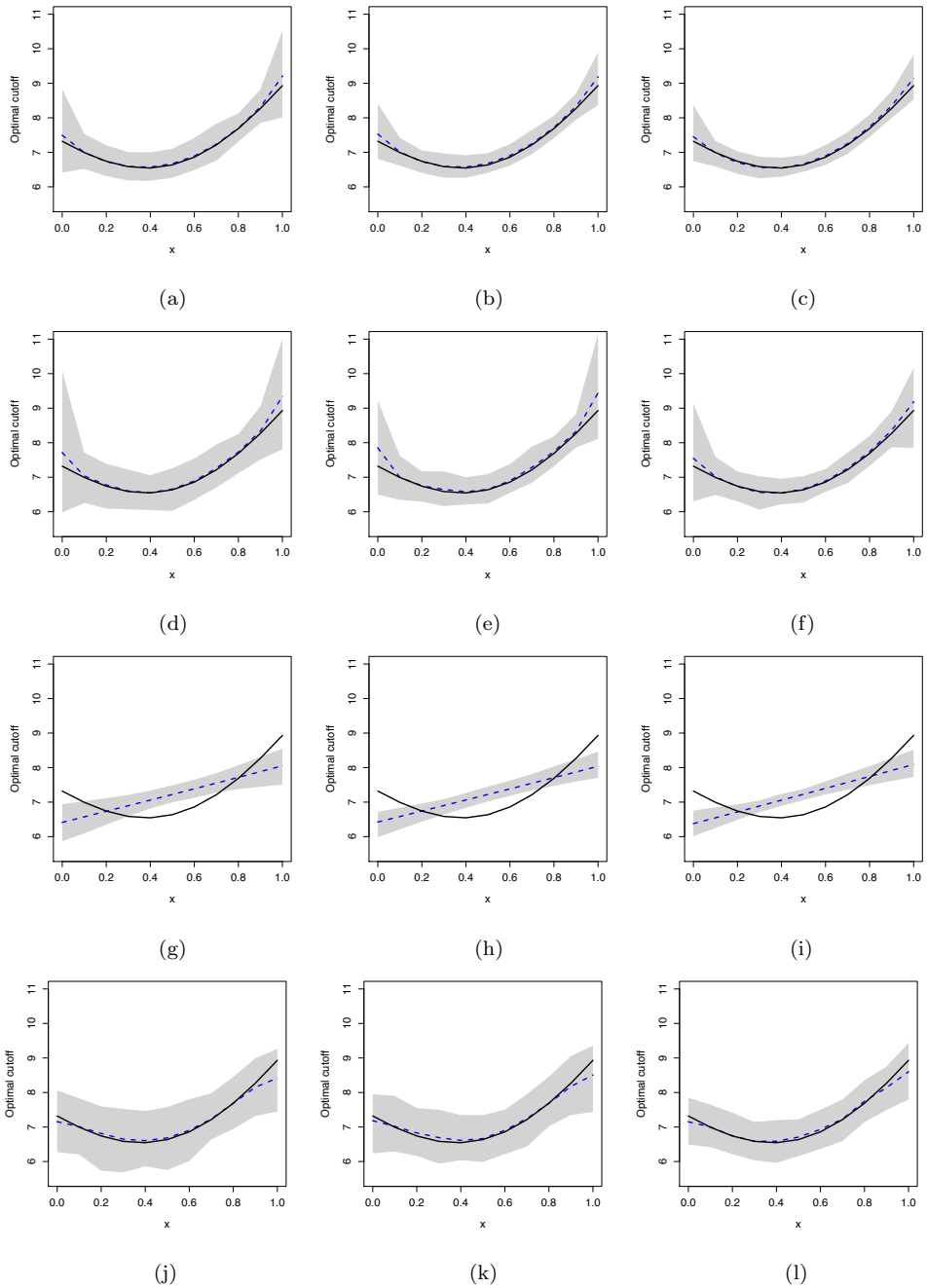


Figure 6. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function under Scenario 3. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator.Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

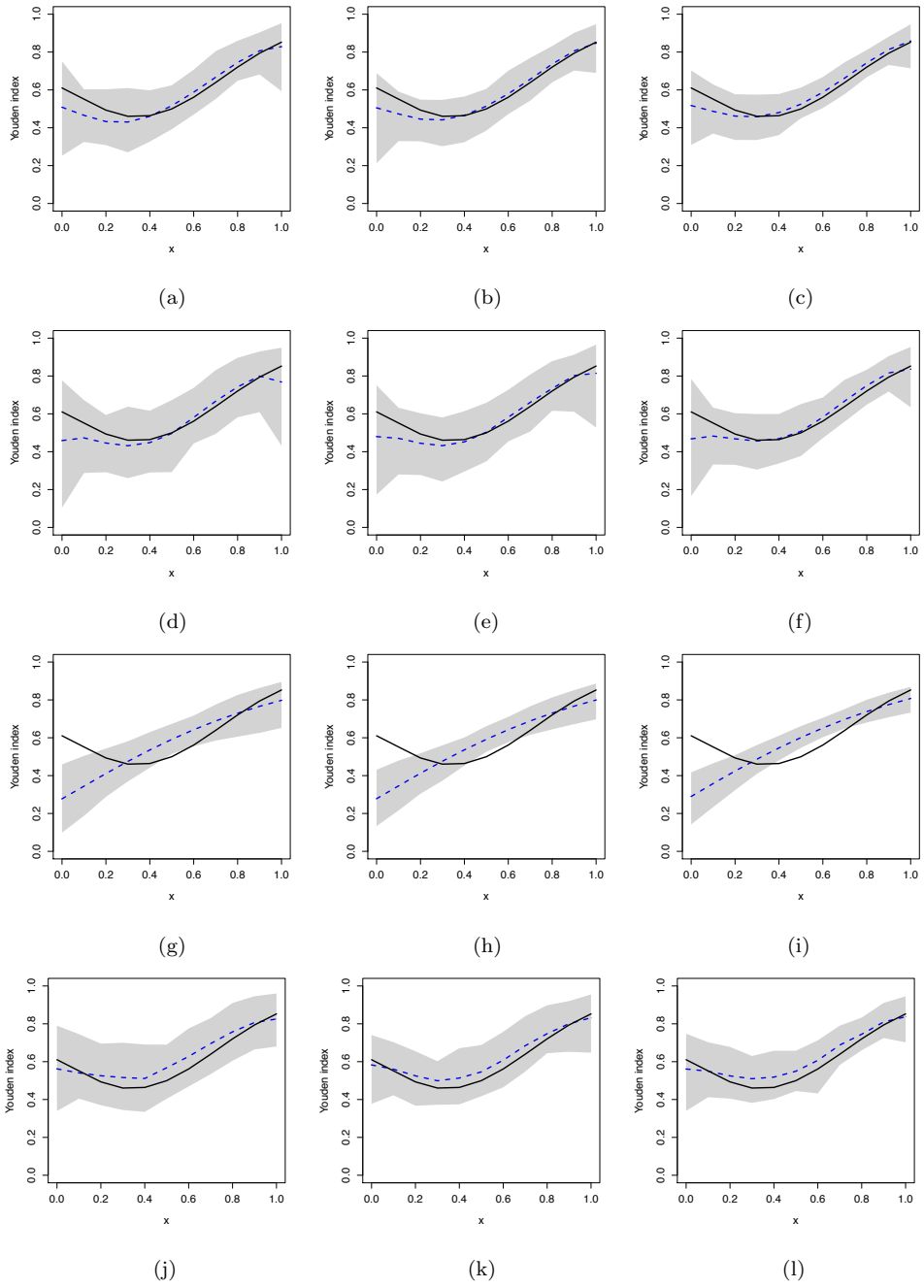


Figure 7. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the Youden index function under Scenario 4. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

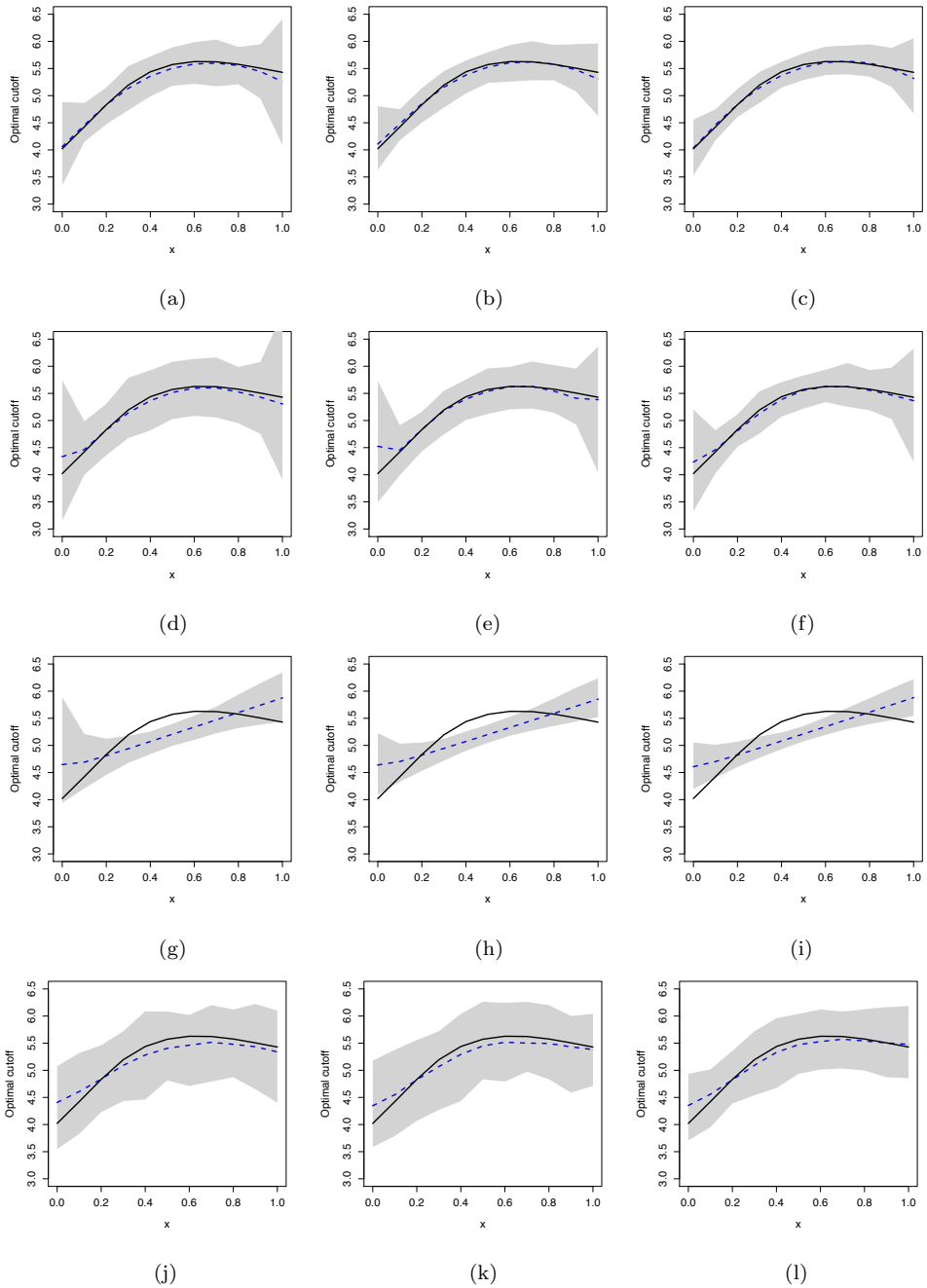


Figure 8. True (solid black line) and mean across simulations (dashed blue line) of the posterior mean (for the Bayesian estimators) of the optimal cutoff function under Scenario 4. A band constructed using the pointwise 2.5% and 97.5% quantiles across simulations is presented in gray. Row 1: B-splines DDP estimator with $Q = 4$. Row 2: B-splines DDP estimator with $Q = 7$. Row 3: Normal estimator. Row 4: Kernel estimator. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

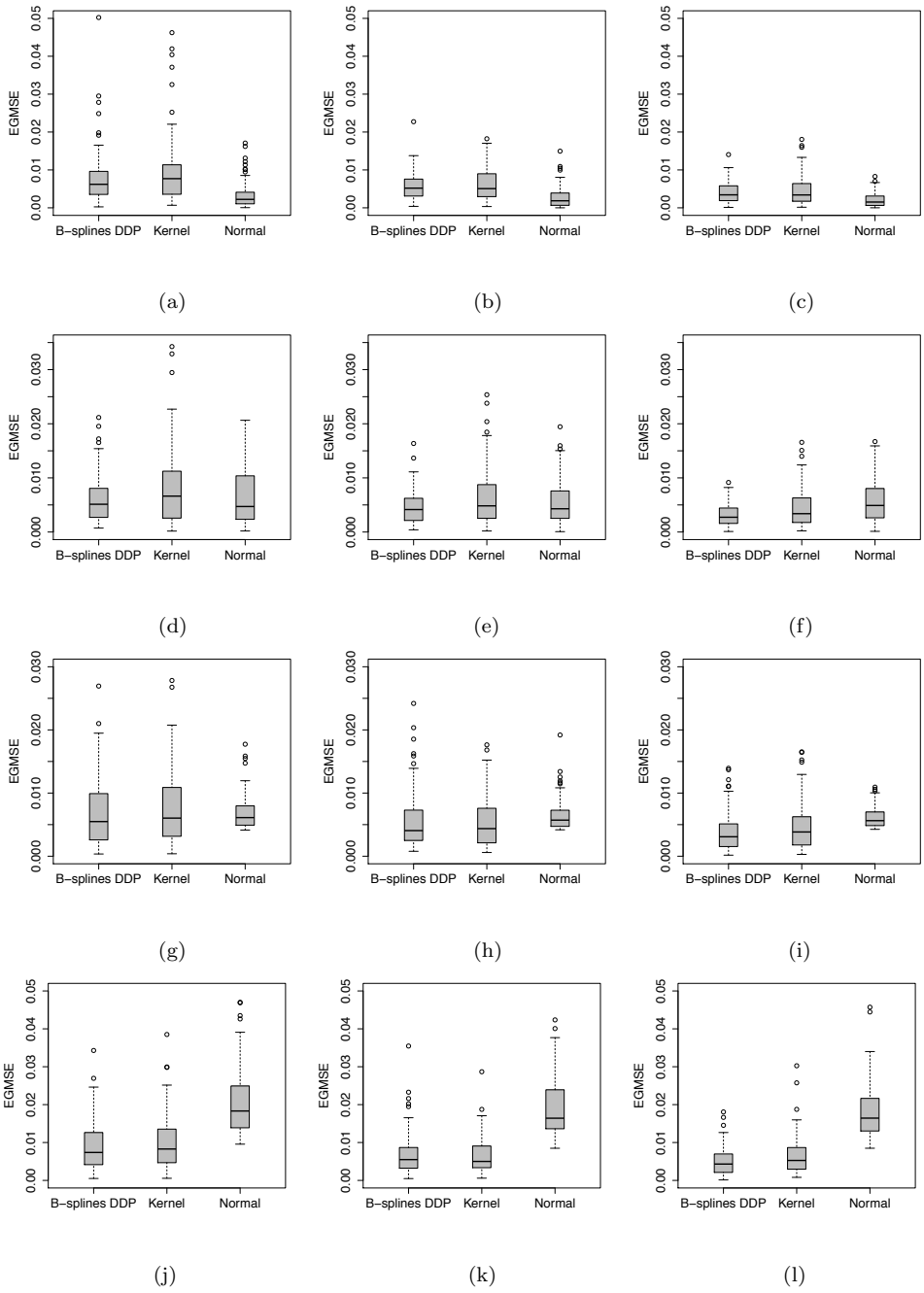


Figure 9. Boxplots of the empirical global mean squared error (EGMSE) of the Youden index across simulations for the B-splines DDP estimator ($Q = 4$), kernel estimator, and normal estimator. Panels (a)–(c), (d)–(f), (g)–(i), and (j)–(l) display the results under Scenario 1, 2, 3, and 4, respectively. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.

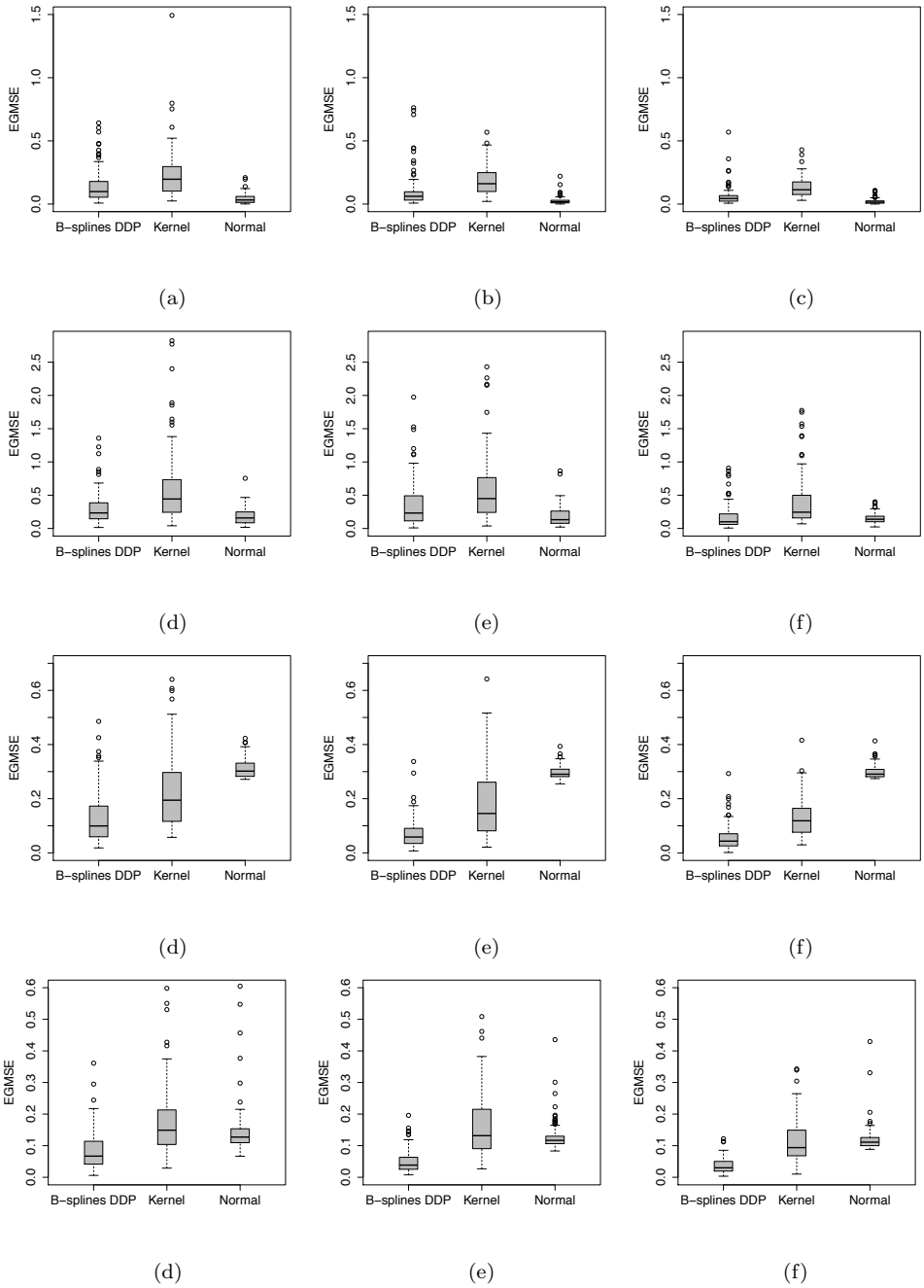
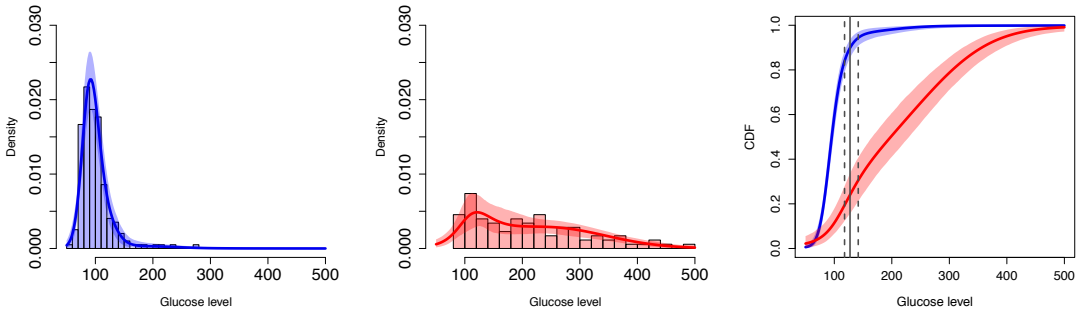
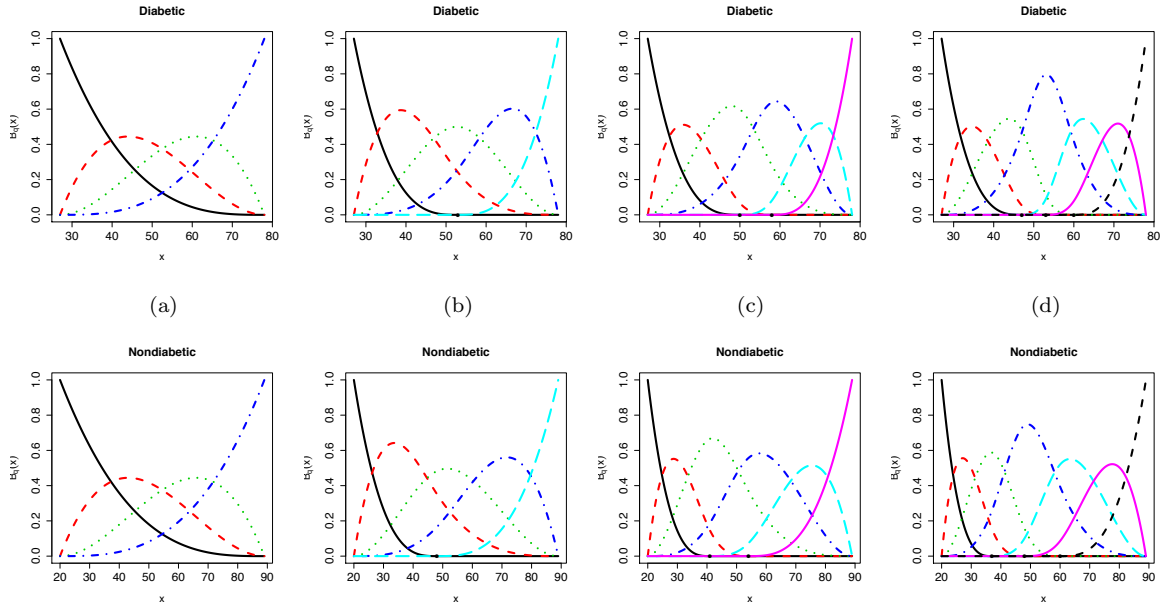


Figure 10. Boxplots of the empirical global mean squared error (EGMSE) of the optimal cutoff across simulations for the B-splines DDP estimator ($Q = 4$), kernel estimator, and normal estimator. Panels (a)–(c), (d)–(f), (g)–(i), and (j)–(l) display the results under Scenario 1, 2, 3, and 4, respectively. Panels (a), (d), (g), and (j) show the results for $(n_D, n_{\bar{D}}) = (100, 100)$, panels (b), (e), (h), and (k) for $(n_D, n_{\bar{D}}) = (100, 200)$, and panels (c), (f), (i), and (l) for $(n_D, n_{\bar{D}}) = (200, 200)$.



(a) (b) (c)

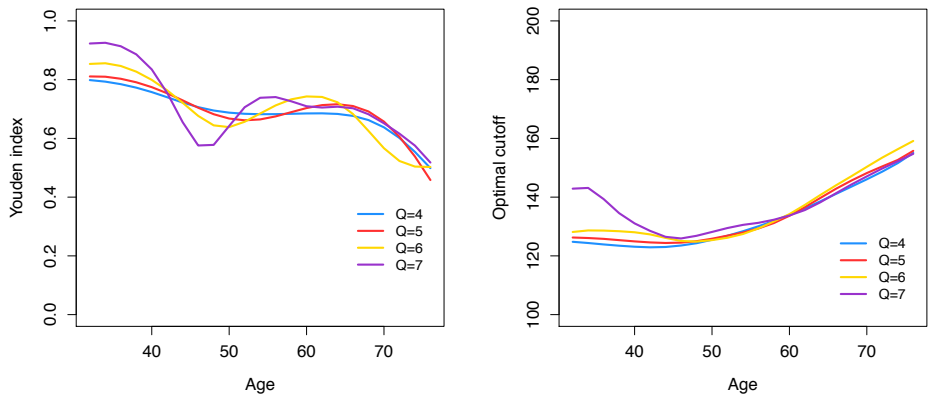
Figure 11. Histogram of the glucose levels in the non-diabetic (panel a) and diabetic group (panel b) along with the posterior mean and 95% pointwise probability interval of the density for each group under (independent) Dirichlet process mixtures of normals. Panel (c) presents the estimated distribution functions and optimal cutoff with its 95% probability interval.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 12. Basis functions for the diabetes data. Panels (a) and (e): $Q = 4$. Panels (b) and (f): $Q = 5$. Panels (c) and (g): $Q = 6$. Panels (d) and (h): $Q = 7$.



(a) (b)

Figure 13. Diabetes data: posterior means of the Youden index (a) and optimal cutoff (b) for the different values of Q considered.

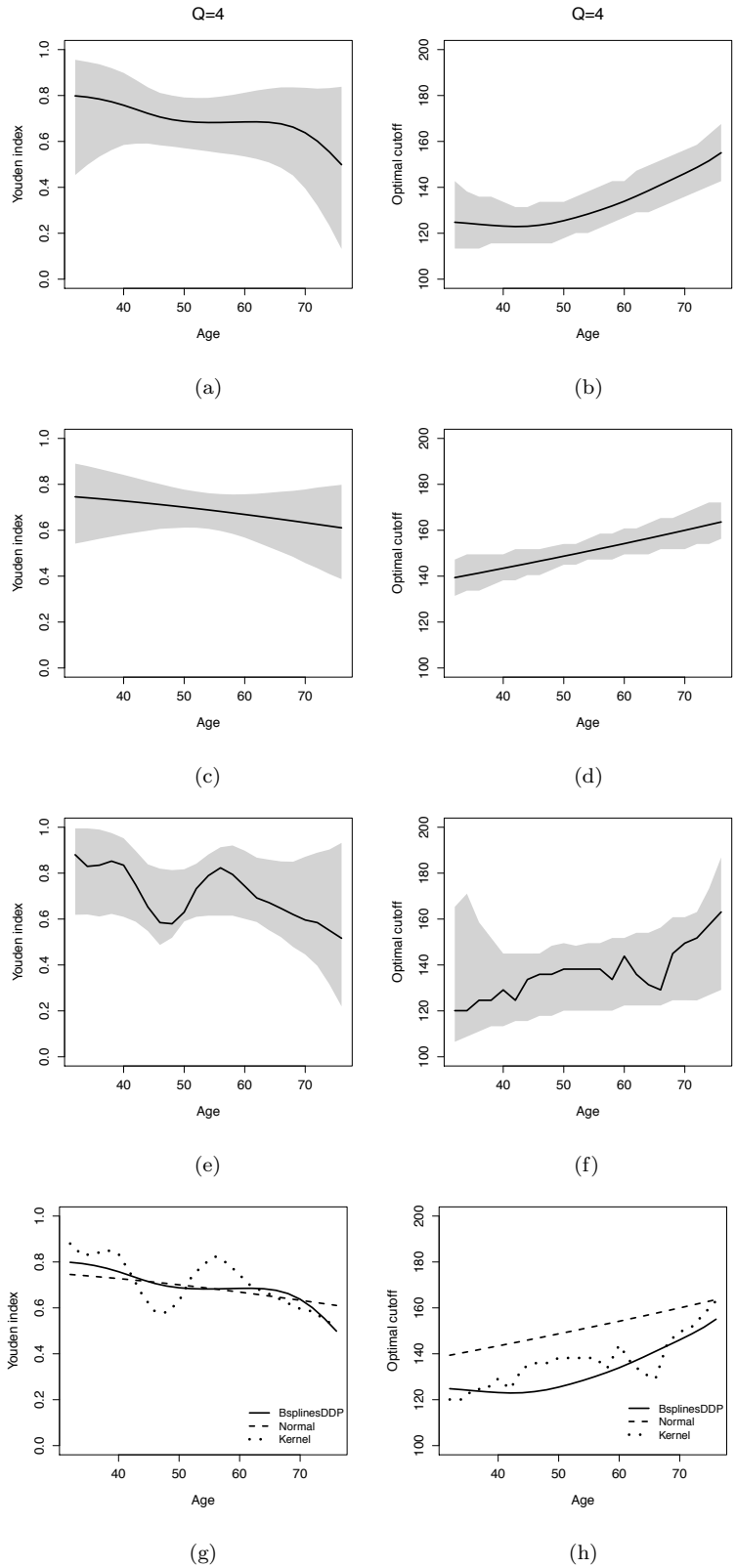


Figure 14. Estimated Youden index and optimal cutoff as a function of age. Panels (a) and (b) present results from the B-splines DDP estimator ($Q = 4$), panels (c) and (d) present results obtained under the normal linear model, while panels (e) and (f) show the results obtained under the nonparametric kernel model. For ease of comparison, panels (g) and (h) display the three estimators together.

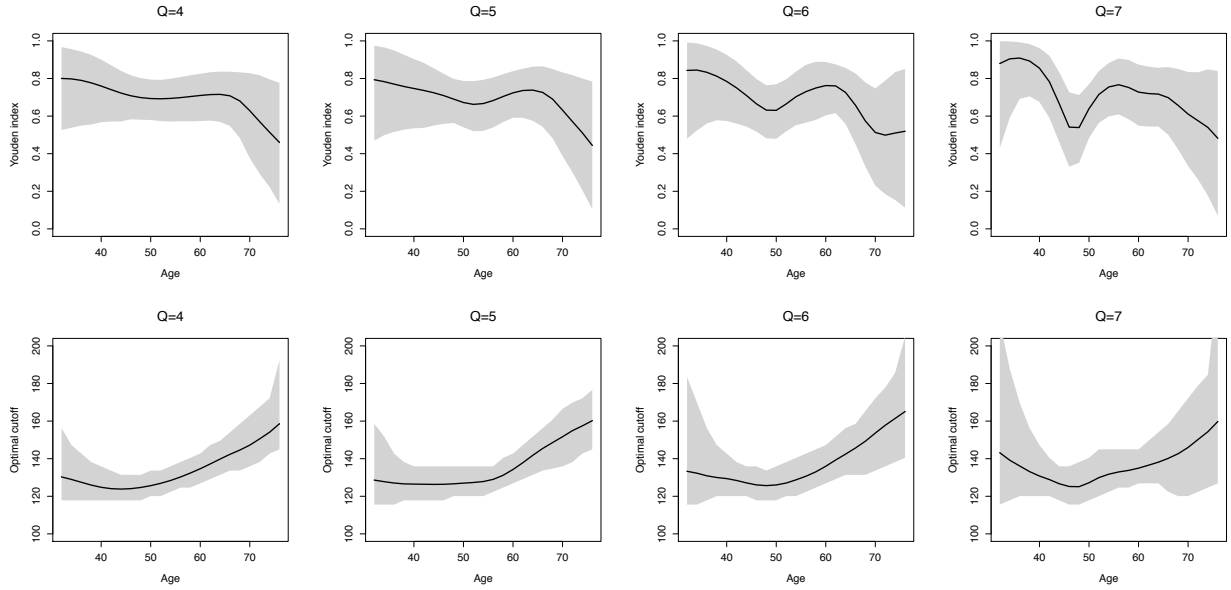


Figure 15. Estimated Youden index and optimal cutoff as a function of age, using the data-driven prior (Section D), for $Q = 4$, $Q = 5$, $Q = 6$, and $Q = 7$. Solid lines represent posterior means and the gray areas correspond to pointwise 95% posterior bands.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	$\text{LPML}_{\bar{D}}$	LPML_{D}	$\text{LPML}_{\bar{D}}$	LPML_{D}	$\text{LPML}_{\bar{D}}$	LPML_{D}	$\text{LPML}_{\bar{D}}$	LPML_{D}
$\mathbf{n_0 = n_1 = 100}$								
B-splines DDP	-187 $(-197, -174)$	-215 $(-224, -204)$	-224 $(-236, -213)$	-212 $(-224, -201)$	-187 $(-198, -174)$	-237 $(-246, -225)$	-178 $(-190, -165)$	-187 $(-197, -176)$
Normal	-185 $(-196, -173)$	-213 $(-222, -202)$	-235 $(-243, -227)$	-220 $(-227, -213)$	-193 $(-203, -180)$	-235 $(-244, -225)$	-182 $(-195, -170)$	-185 $(-195, -175)$
$\mathbf{n_0 = n_1 = 200}$								
B-splines DDP	-368 $(-383, -351)$	-427 $(-441, -411)$	-440 $(-452, -427)$	-414 $(-428, -399)$	-368 $(-382, -351)$	-471 $(-485, -456)$	-351 $(-368, -336)$	-371 $(-384, -356)$
Normal	-366 $(-382, -350)$	-425 $(-439, -409)$	-471 $(-480, -460)$	-436 $(-446, -424)$	-381 $(-395, -362)$	-470 $(-483, -453)$	-360 $(-376, -343)$	-369 $(-384, -353)$

Table 1

Average LPML and 90% *interval for each model in each simulation scenario.*

	Prior 1		Prior 2	
	$\text{LPML}_{\bar{D}}$	LPML_D	$\text{LPML}_{\bar{D}}$	LPML_D
$Q = 4$	-878	-530	-880	-530
$Q = 5$	-880	-532	-881	-531
$Q = 6$	-881	-531	-883	-531
$Q = 7$	-887	-532	-886	-532

Table 2

Diabetes data: LPML for the different values of Q considered under prior configuration 1 (the one used on the main manuscript) and under prior configuration 2 (the one described in Section D of this Supplementary Material).